bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Submitted to the Annals of Applied Statistics

A NOVEL AND EFFICIENT ALGORITHM FOR DE NOVO DISCOVERY OF MUTATED DRIVER PATHWAYS IN CANCER

By Binghui Liu[†,‡], Chong Wu[‡], Xiaotong Shen[‡] and Wei Pan[*,‡]

Northeast Normal University[†] *and University of Minnesota*[‡]

Next-generation sequencing studies on cancer somatic mutations have discovered that driver mutations tend to appear in most tumor samples, but they barely overlap in any single tumor sample, presumably because a single driver mutation can perturb the whole pathway. Based on the corresponding new concepts of coverage and mutual exclusivity, new methods can be designed for de novo discovery of mutated driver pathways in cancer. Since the computational problem is a combinatorial optimization with an objective function involving a discontinuous indicator function in high dimension, many existing optimization algorithms, such as a brute force enumeration, gradient descent and Newton's methods, are practically infeasible or directly inapplicable. We develop a new algorithm based on a novel formulation of the problem as non-convex programming and non-convex regularization. The method is computationally more efficient, effective and scalable than existing Monte Carlo searching and several other algorithms, which have been applied to The Cancer Genome Atlas (TCGA) project. We also extend the new method for integrative analysis of both mutation and gene expression data. We demonstrate the promising performance of the new methods with applications to three cancer datasets to discover de novo mutated driver pathways.

1. Introduction. It is known that cancer is characterized by numerous somatic mutations, of which only a subset, named *"driver"* mutations, contribute to tumor growth and progression. With next-generation whole-genome or whole-exome sequencing, somatic mutations are measured in large numbers of cancer samples (Mardis & Wilson, 2008; Meyerson *et al.*, 2010). To improve understanding and treatment of cancers, it is critical to distinguish driver mutations from neutral *"passenger"* mutations. A standard approach to predicting driver mutations is to identify recurrent mutations in cancer patients (Beroukhim *et al.*, 2007; Getz *et al.*, 2007), which has its drawback in its inability to capture mutational heterogeneity of cancer genomes (Ding *et al.*, 2008; Jones *et al.*, 2008). An emerging discovery is that

*Corresponding author.

Keywords and phrases: DNA sequencing, Driver mutations, Optimization, Subset selection, Truncated L_1 penalty

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a given sample driver mutations typically target one, but not all, of several genes in cellular signaling and regulatory pathways (Vogelstein & Kinzler, 2004). Hence the research has shifted from the gene level to pathway level (Boca, 2010; Efroni, 2011). Recent studies indicated that mutations arising in driver pathways often cover a majority of samples, but, importantly, for a single sample only a single or few mutations appear because a single mutation is capable to perturb the whole pathway; the latter concept is the so-called *mutual exclusivity*. By using mutual exclusivity, new pathway-based methods are developed to identify de novo driver mutations and pathways (Ciriello *et al.*, 2012; Masica *et al.*, 2011; Miller *et al.*, 2011). For example, Miller *et al.* (2011) proposed a method to find functional sets of mutations by using patterns of recurrent and mutually exclusive aberrations; Ciriello *et al.* (2012) not only used the mutual exclusivity pattern, but also incorporated a gene functional network constructed based on prior knowledge. Recently Vandin *et al.* (2012) introduced a novel scoring function combining the two concepts, **coverage** and **mutual exclusivity**, to identify mutated driver pathways through optimizing this scoring function, which has been used in some large-scale cancer sequencing studies. It is solved by stochastic search methods: a greedy algorithm and a Markov chain Monte Carlo method. Other proposals based on binary linear programming, genetic search algorithm, and integer linear programming have appeared (Zhao *et al.*, 2012; Leiserson *et al.*, 2013), all of which are still relatively slow, especially for large-scale problems. To address these issues, we reformulate the problem of identifying mutated driver pathways as a statistical problem of subset identification to minimize a new cost function, what we call minimum cost subset selection (MCSS). A key component is a novel approximation to a combinatorial problem through regularization, where a discontinuous indicator function is approximated by a continuous and non-convex truncated L_1 (TL) function (Shen *et al.*, 2012). Furthermore, we add a truncated L_1 penalty (TLP) to the cost function to seek a sparse solution, as well as adding a small ridge penalty to alleviate the problem of multiple solutions. As a result, a combinatorial optimization problem becomes a continuous but non-convex one in the Euclidean space, which can be efficiently solved through a non-convex optimization technique, leading to high computational improvement.

Another advantage of the proposed method is that it is able to find multiple mutated driver pathways. An existing method to identify multiple mutated driver pathways is Multi-Dendrix (Leiserson *et al.*, 2013), in which the number of pathways and the number of the genes in each pathway have to be specified in advance. On the contrary, our proposed method does not

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

need to fix such numbers beforehand. Based on a series of randomly selected initial estimates, a series of low-cost estimates of mutated driver pathways can be obtained. Moreover, the proposed method is general so that other types of information can be incorporated in a simple way. For example, if a gene interaction network is available, it can be incorporated by adding a network-based penalty to the current cost function as in Li & Li (2008); since it is more informative to combine mutation data with other types of data such as gene expression data (Zhang and Zhou , 2014), an integrative version can be developed by adding other cost functions for other types of data into the current one. As a concrete example, we propose a new method to integrate mutation data with gene expression data.

2. Methods.

2.1. *Problem.* Consider mutation data with n patients and p genes, represented as an $n \times p$ mutation matrix $\boldsymbol{A}$ with entry $A_{ij} = 1$ if gene j is mutated in patient i, and $A_{ij} = 0$ otherwise. For gene $j \in V = \{1, ..., p\}$, let $\Gamma(j) = \{i : A_{ij} = 1\}$ be a subgroup of patients whose gene j is mutated. Moreover, given a subset of genes $B \subseteq \{1, ..., p\}$, let $\Gamma(B)$ be a subgroup of patients with at least one of the genes in B mutated, i.e. $\Gamma(B) = \bigcup_{j \in B} \Gamma(j)$. Cancer sequencing studies have motivated to identify a set of mutated genes across a large number of patients, whereas only a small number of patients have mutations in more than one gene in the set, that is, these mutations are approximately exclusive. This amounts to finding a set $B \subseteq V$ of genes such that (i) the coverage is high, that is, most patients have at least one mutation in B; (ii) the genes in B are approximately exclusive, that is, most patients have no more than one mutation in B. A measure $\omega(B) = \sum_{j \in B} |\Gamma(j)| - |\Gamma(B)|$ was proposed by Vandin *et al.* (2012), called the coverage overlap, to balance the trade-off between coverage and exclusivity. To maximize the coverage $|\Gamma(B)|$ and minimize the coverage overlap $\omega(B)$ simultaneously, Vandin *et al.* (2012) suggests to minimize

$$f(B) = \frac{\omega(B)}{n} - \frac{|\Gamma(B)|}{n} = \frac{1}{n} \sum_{j \in B} |\Gamma(j)| - \frac{2}{n} |\Gamma(B)| \tag{2.1}$$

with respect to B, thus obtaining an estimate $\hat{B}$. Minimizing $f(B)$ is equivalent to maximizing the weight function $-f(B)$, which is called the maximum weight sub-matrix problem (MWSP). Note that minimizing $f(B)$ is a non-trivial combinatorial problem, to which most existing optimization algorithms based on the gradient descent or Newton's algorithm cannot be directly applied. A popular method called Dendrix is based on a Monte

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Carlo search algorithm to seek an approximate solution to minimize $f(B)$ (Vandin *et al.*, 2012).

2.2. *New formulation.* As indicated in (2.1), MWSP is a combinatorial problem, for which a brute force search is time-consuming and not scalable for large (n, p), while many existing algorithms like gradient descent or Newton's method cannot be directly applied. Here we formulate it as nonconvex minimization and examine a regularized version by imposing penalties to ensure proper solutions. Specifically, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, let $B = B(\boldsymbol{\beta}) = \{j \in V : |\beta_j| \neq 0\}$, and we rewrite $|\Gamma(B)| = \sum_{i=1}^n I\big(\sum_{j=1}^p A_{ij} I(|\beta_j| \neq 0) \neq 0\big)$, $\sum_{j\in B} |\Gamma(j)| = \sum_{j=1}^p I(|\beta_j| \neq 0) A_{\cdot,j}$, $A_{\cdot,j} = \sum_{i=1}^n A_{ij}$ and $|\Gamma(j)| = A_{\cdot,j}$ for each $j \in \{1, \cdots, p\}$. Then (2.1) becomes

$$f(B(\boldsymbol{\beta})) = \frac{1}{n} \sum_{j=1}^{p} I(|\beta_j| \neq 0) A_{\cdot,j} - \frac{2}{n} \sum_{i=1}^{n} I\Big(\sum_{j=1}^{p} A_{ij} I(|\beta_j| \neq 0) \neq 0\Big). \tag{2.2}$$

Minimizing (2.2) in $\boldsymbol{\beta}$ yields an estimate $\check{\boldsymbol{\beta}} = (\check{\beta}_1, \cdots, \check{\beta}_p)'$, and thus an estimated set $\check{B} = \{j : |\check{\beta}_j| \neq 0\}$. However, due to the discontinuity with the indicator function $I(.)$, it is difficult to minimize (2.2) directly; instead, since $\min(|\beta_j|/\tau_1, 1) \to I(|\beta_j| \neq 0)$ as $\tau_1 \to 0^+$, we propose a surrogate to minimize

$$\begin{aligned} S(\boldsymbol{\beta}) \;=\; & \frac{1}{n} \sum_{j=1}^{p} \min\left(\beta_j/\tau_1, 1\right) A_{\cdot,j} - \frac{2}{n} \sum_{i=1}^{n} \min\Big(\sum_{j=1}^{p} A_{ij}\beta_j/\tau_1, 1\Big) \\ & + \lambda \sum_{j=1}^{p} \min\left(\beta_j/\tau_2, 1\right) + \frac{\alpha}{n} \sum_{j=1}^{p} \beta_j^2, \end{aligned} \tag{2.3}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in [0, +\infty)^p$; that is, $\boldsymbol{\beta}$ is a vector of parameters to be estimated; λ, α, τ_1 and τ_2 are non-negative tuning parameters to be determined via a grid search in cross-validation (as used in the later experiments); A_{ij}'s are observed and known. Note that in (2.3), the last two terms, as a TLP and a ridge penalty respectively, ensure sparse and proper solutions.

2.3. *Computation.* To solve nonconvex minimization (2.3), we employ difference convex (DC) programming by decomposing the objective function into a difference of two convex functions, on which convex relaxation is performed through iterative approximations of the trailing convex function through majorization. Specifically, $\min(\frac{z}{\tau}, 1)$ can be written as a difference of two convex functions: $\min(z/\tau, 1) = z/\tau - \max\left(z/\tau - 1, 0\right)$ for any $z > 0$

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and $\tau > 0$. Then, we obtain a sequence of upper approximations $S^{(m)}(\boldsymbol{\beta})$ of $S(\boldsymbol{\beta})$ at iteration m (up to a constant) as follows:

$$S^{(m)}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\big(\mathrm{diag}(\boldsymbol{A}_{\cdot})I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_1)/n\tau_1 + \lambda I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_2)/\tau_2 - \frac{2\boldsymbol{A}_{\cdot}}{n\tau_1}\big)$$

$$+ \frac{2}{n}\sum_{i=1}^{n}\max\Big(\sum_{j=1}^{p} A_{ij}\beta_j/\tau_1 - 1, 0\Big) + \frac{\alpha}{n}\boldsymbol{\beta}'\boldsymbol{\beta}, \tag{2.4}$$

where $\boldsymbol{\beta} \in [0, +\infty)^p$, $\boldsymbol{A}_{\cdot} = (A_{\cdot,1}, \ldots, A_{\cdot,p})'$, and $\mathrm{diag}(\boldsymbol{A}_{\cdot})$ is a diagonal matrix with elements of $\boldsymbol{A}_{\cdot}$ as diagonals. Now $S^{(m)}(\boldsymbol{\beta})$ is strictly convex (since the first term is linear in $\boldsymbol{\beta}$, the second is convex while the last is quadratic in $\boldsymbol{\beta}$ with $\alpha \neq 0$), we use some existing convex program package (CVX in Matlab), or more efficiently, the subgradient descent method (as shown in the appendix) (Shor , 1985), to obtain a unique minimizer $\hat{\boldsymbol{\beta}}^{(m)}$; we repeat the process until convergence to obtain $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(+\infty)}$.

Interestingly, one may replace the TLP in $S(\boldsymbol{\beta})$ in (2.3) with the L_1-penalty, yielding $\hat{\boldsymbol{\beta}}^{L}$. This, together with, other randomly generated numbers, can be use as an initial value $\hat{\boldsymbol{\beta}}^{(0)}$ for our method. For selection of tuning parameters, we may consider cross-validation, as discussed later.

The following algorithm summarizes our computational method.

ALGORITHM 1. *Given the parameters τ_1, τ_2, λ, α.*

Initialization *Supply an initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$.*

Iteration *At iteration m, compute $\hat{\boldsymbol{\beta}}^{(m)}$ by minimizing (2.4).*

Stopping rule *Terminate when $S(\hat{\boldsymbol{\beta}}^{(m-1)}) - S(\hat{\boldsymbol{\beta}}^{(m)}) \le 0$. The estimate is $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(m^\star-1)}$, where $m^\star$ is the smallest index satisfying the termination criterion. The estimated subset is $\tilde{B} = \{j \in \{1, \cdots, p\} : \tilde{\beta}_j \neq 0\}$.*

The following convergence property of Algorithm 1 has been established.

THEOREM 1. *$\hat{\boldsymbol{\beta}}^{(m)}$ in Algorithm 1 converges in finite steps to a local minimizer $\tilde{\boldsymbol{\beta}}$ of $S(\boldsymbol{\beta})$ in (2.3). $S(\hat{\boldsymbol{\beta}}^{(m)})$ strictly decreases in m until $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} = \hat{\boldsymbol{\beta}}^{(m^\star-1)}$ for all $m \ge m^\star$.*

2.4. *Initial estimate.* In general, a large number of good or randomly selected initial estimates may be used to obtain multiple solutions, from which a subset of more promising ones with smaller objective or cost function

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

values can be selected. Below, we describe a simple way to obtain a good initial estimate, which was used in later simulations; we modify $S(\boldsymbol{\beta})$ such that the modified version $S_L(\boldsymbol{\beta})$ becomes much easier to optimize. A local condition of (2.3) can be established based on regular subdifferentials

$$\frac{A_{\cdot,j}b_j^{(1)}}{n\tau_1} + 2\sum_{i=1}^{n}\frac{b_{ij}^{(2)}}{n\tau_1} + \frac{\lambda b_j^{(3)}}{\tau_2} + 2\frac{\alpha}{n}\beta_j = 0, \quad j = 1, \cdots, p,$$

where $b_j^{(1)} \in [-1,1]$ if $\beta_j = 0$, $b_j^{(1)} = \text{sign}(\beta_j)$ if $0 < |\beta_j| < \tau_1$, $b_j^{(1)} = 0$ if $|\beta_j| > \tau_1$ and $b_j^{(1)} = \emptyset$ if $|\beta_j| = \tau_1$ for $j = 1, \cdots, p$; $b_j^{(3)} \in [-1,1]$ if $\beta_j = 0$, $b_j^{(3)} = \text{sign}(\beta_j)$ if $0 < |\beta_j| < \tau_2$, $b_j^{(1)} = 0$ if $|\beta_j| > \tau_2$ and $b_j^{(3)} = \emptyset$ if $|\beta_j| = \tau_2$ for $j = 1, \cdots, p$. Note that $b_{ij}^{(2)}$ is more complicated as it depends on the values of $A_{ij'}$ and $\beta_{j'}$, $j' \in \{1, \cdots, p\}$, and $b_{ij}^{(2)} = 0$ or $b_{ij}^{(2)} = -A_{ij}$ or $b_{ij}^{(2)} \in [-A_{ij}, 0]$ for $\beta_j > 0$. Based on these regular subdifferentials, we develop the following lemma.

LEMMA 1. *If there exists a non-zero local minimizer $\boldsymbol{\beta}^*$ of $S(\boldsymbol{\beta})$ in (2.3) on $\mathbb{R}^p$, then $0 \le |\beta_j^*| \le \tau_1$ for each $j \in \{1, \cdots, p\}$.*

Lemma 1 says that the set of all local minimizers of $S(\boldsymbol{\beta})$ in (2.3) over $[0, +\infty]^p$ is the same as that obtained from the following cost function over $[0, \tau_1]^p$:

$$\begin{aligned} S(\boldsymbol{\beta}) \;=\; & \frac{1}{n}\sum_{j=1}^{p}\frac{\beta_j A_{\cdot,j}}{\tau_1} - \frac{2}{n}\sum_{i=1}^{n}\min\Big(\frac{\sum_{j=1}^{p}A_{ij}\beta_j}{\tau_1}, 1\Big) \\ & + \frac{\alpha}{n}\sum_{j=1}^{p}\beta_j^2 + \lambda\sum_{j=1}^{p}\min(\frac{\beta_j}{\tau_2}, 1), \qquad \boldsymbol{\beta} \in [0, \tau_1]^p. \end{aligned} \tag{2.5}$$

If we use the L_1-penalty as opposed to the truncated L_1-penalty in (2.5), then the cost function becomes

$$\begin{aligned} S_L(\boldsymbol{\beta}) = & \frac{1}{n}\sum_{j=1}^{p}\frac{\beta_j A_{\cdot,j}}{\tau_1} - \frac{2}{n}\sum_{i=1}^{n}\min\Big(\frac{\sum_{j=1}^{p}A_{ij}\beta_j}{\tau_1}, 1\Big) \\ & + \frac{\alpha}{n}\sum_{j=1}^{p}\beta_j^2 + \lambda\sum_{j=1}^{p}\frac{\beta_j}{\tau_2}, \qquad \boldsymbol{\beta} \in [0, \tau_1]^p, \end{aligned}$$

which is strictly convex in $\boldsymbol{\beta} \in [0, \tau_1]^p$, yielding a unique minimizer $\hat{\boldsymbol{\beta}}_L$.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.5. *Model selection.* Tuning parameters (λ, r) need to be estimated from data, where $\tau_2 = r\tau_1$ $(0 < r < 1)$, while α is fixed at a sufficiently small positive number, say $\alpha = 10^{-3}$, and τ_1 is fixed at any positive value, say $\tau_1 = 1$. Tuning of (λ, r) can be achieved through sample splitting. As a matter of fact, the term $\frac{\alpha}{n}\sum_{j=1}^{p}\beta_j^2$ is introduced to yield a unique minimizer of (2.4) so that the bias caused by the ridge penalty is ignorable for sufficiently small α. On the other hand, given the ratio r, an exact value of (τ_1, τ_2) is unimportant. This is because $\min_{\boldsymbol{\beta}\in[0,+\infty)^p} S(\boldsymbol{\beta}; K\tau_1, r, \lambda, \alpha') = \min_{\boldsymbol{\beta}'\in[0,+\infty)^p} S(\boldsymbol{\beta}'; \tau_1, r, \lambda, \alpha') = \min_{\boldsymbol{\beta}\in[0,+\infty)^p} S(\boldsymbol{\beta}; \tau_1, r, \lambda, \alpha')$ if $S(\boldsymbol{\beta}; K\tau_1, r, \lambda, \alpha') = S(\boldsymbol{\beta}'; \tau_1, r, \lambda, \alpha')$ with $\boldsymbol{\beta}' = \boldsymbol{\beta}/K$ and $\alpha = \alpha'/\tau_1^2$ for any $K > 0$. Consequently, given the ratio r, optimization in terms of different choices of τ_1 are equivalent. Given a $n \times p$ mutation matrix $\boldsymbol{A}$, a candidate set $\Lambda \subseteq (0, +\infty)$ of the tuning parameter λ and a candidate set $R \subseteq (0, +\infty)$ of the tuning parameter $r = \tau_2/\tau_1$, we use a sample splitting procedure to select the tuning parameters $\hat{\lambda} \in \Lambda$ and $\hat{r} \in R$:

Initialization Supply a randomly selected initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$.

Partition Randomly partition the rows of the mutation matrix A into two parts: training data $\boldsymbol{A}^{tr}$ and tuning data $\boldsymbol{A}^{tu}$.

Training For each $\lambda \in \Lambda$ and each $r \in R$, apply Algorithm 1 to the training data $\boldsymbol{A}^{tr}$ with the initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$ and parameters λ and r to get the corresponding estimate $\hat{\boldsymbol{\beta}}^{tr}(\lambda, r)$.

Tuning Based on the tuning data $\boldsymbol{A}^{tu}$, we formulate a tuning error for each $\hat{\boldsymbol{\beta}}^{tr}(\lambda, r)$ as

$$\begin{aligned}\mathrm{TE}(\hat{\boldsymbol{\beta}}^{tr}(\lambda, r), \boldsymbol{A}^{tu}) =& \frac{1}{n^{tu}}\sum_{j=1}^{p} I(\hat{\beta}^{tr}(\lambda, r)_j > 0) A^{tu}_{\cdot,j} \\ &- \frac{2}{n^{tu}}\sum_{i=1}^{n^{tu}} I\big(\sum_{j=1}^{p} A^{tu}_{ij}\hat{\beta}^{tr}(\lambda, r)_j > 0\big),\end{aligned}$$

where n^{tu} denotes the number of rows of $\boldsymbol{A}^{tu}$, that is, the patient number in the tuning data, and $A^{tu}_{\cdot,j} = \sum_{i=1}^{n^{tu}} A^{tu}_{ij}$. We select λ and r as

$$(\hat{\lambda}, \hat{r}) = \arg\min_{(\lambda,r)\in\Lambda\times R} \mathrm{TE}(\hat{\boldsymbol{\beta}}^{tr}(\lambda, r), \boldsymbol{A}^{tu}).$$

Given $\lambda = \hat{\lambda}$ and $r = \hat{r}$, we apply Algorithm 1 to the original mutation matrix $\boldsymbol{A}$ to find $\hat{\boldsymbol{\beta}} \in [0, +\infty)^p$ that minimizes $S(\boldsymbol{\beta})$ in (2.3).

2.6. *Integrative analysis.* An advantage of the proposed algorithm is its possible extensions to include other types of genomic data, in addition to

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mutation data. To this end, we modify the proposed cost function and algorithm to incorporate other types of data such as gene expression. Let $f_{ME}(B)$ denote the integrative cost function, which is the sum of the original cost function $f(B)$ and a new one $f_E(B)$ for gene expression data:

$$f_{ME}(B) = f(B) + \gamma f_E(B) = \frac{1}{n}\sum_{j\in B}|\Gamma(j)| - \frac{2}{n}|\Gamma(B)| - \gamma \sum_{j,k\in B, j\neq k} c_{jk}$$

where c_{jk} is the Pearson correlation coefficient of the expression profiles of genes j and k. Note that the integrative cost function is based on the observation that the genes in the same pathway usually collaborate with each other to execute a common function. Therefore, the expression profiles of the genes in the same pathway usually have higher correlations than those from different pathways (Qiu *et al.*, 2010; Zhao *et al.*, 2012).

To minimize $f_{ME}(B)$, we develop a similar algorithm as before, called MCSS_ME, where $S(\boldsymbol{\beta})$ and $S^{(m)}(\boldsymbol{\beta})$ are replaced by $S_{ME}(\boldsymbol{\beta})$ and $S^{(m)}_{ME}(\boldsymbol{\beta})$ respectively as follows.

$$\begin{aligned} S_{ME}(\boldsymbol{\beta}) =& \frac{1}{n}\sum_{j=1}^{p}\min\left(\beta_j/\tau_1, 1\right) A_{\cdot,j} - \frac{2}{n}\sum_{i=1}^{n}\min\left(\sum_{j=1}^{p} A_{ij}\beta_j/\tau_1, 1\right) \\ & - \gamma\sum_{j,k} c_{jk}\min\left(\beta_j/\tau_1, 1\right)\min\left(\beta_k/\tau_1, 1\right) \\ & + \lambda\sum_{j=1}^{p}\min\left(\beta_j/\tau_2, 1\right) + \frac{\alpha}{n}\sum_{j=1}^{p}\beta_j^2 \end{aligned}$$

and

$$\begin{aligned} S^{(m)}_{ME}(\boldsymbol{\beta}) =& \boldsymbol{\beta}'\Big(\frac{\mathrm{diag}(\boldsymbol{A}_{\cdot})I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_1)}{n\tau_1} + \frac{\lambda I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_2)}{\tau_2} - 2\gamma\frac{\boldsymbol{D}\hat{\boldsymbol{\beta}}^{(m-1)}}{\tau_1^2} \\ & - 2\frac{\boldsymbol{A}_{\cdot}}{n\tau_1} - 2\gamma\mathrm{diag}(I(\hat{\boldsymbol{\beta}}^{(m-1)} > \tau_1))\boldsymbol{D}\frac{\max(\hat{\boldsymbol{\beta}}^{(m-1)}/\tau_1 - 1, 0)}{\tau_1}\Big) \\ & + 2\gamma\boldsymbol{C}_{\cdot}'\mathrm{diag}(\max(\boldsymbol{\beta}/\tau_1 - 1, 0))\max(\boldsymbol{\beta}/\tau_1 - 1, 0) \\ & + 2\gamma\boldsymbol{C}_{\cdot}'\mathrm{diag}(\boldsymbol{\beta})\boldsymbol{\beta}/\tau_1^2 + 4\gamma\max(\boldsymbol{\beta}/\tau_1 - 1, 0)'\boldsymbol{C}\boldsymbol{\beta}/\tau_1 \\ & + \frac{2}{n}|\max(\boldsymbol{A}\boldsymbol{\beta}/\tau_1 - 1, 0)| + \frac{\alpha}{n}\boldsymbol{\beta}'\boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{D} = \boldsymbol{C} + \mathrm{diag}(\boldsymbol{C}_{\cdot})$, $\boldsymbol{C} = [c_{jk}]$ ($c_{jj} = 0$) and $\boldsymbol{C}_{\cdot}$ is the row sum vector of $\boldsymbol{C}$. Here we use the subgradient descent method (as shown in the appendix) to obtain a minimizer of $S^{(m)}_{ME}(\boldsymbol{\beta})$.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To choose a suitable γ in situations with no prior information, we propose a method to balance the contributions to the new cost function from mutation data and from gene expression data. Specifically, we randomly select a large number of subsets, say B_1, B_2,..., B_R, of the genes from $\{1, 2, \cdots, p\}$ with the size of each subset $|B_j|$ randomly generated from $\{2, 3, \cdots, n_p\}$, then we choose $\gamma = \min_j f(B_j)/\min_j f_E(B_j)$, which aims to give an equal weight on the contribution of the mutation data and that of the expression data to the overall cost function $f_{ME}()$. In our following experiments, we always used $R = 10000$ and $n_p = 8$, though other values may be used.

After determining γ, we choose the other tuning parameters similarly as before but according to an integrative version of the tuning error

$$\begin{aligned}
&\mathrm{TE}_{\mathrm{ME}}(\hat{\boldsymbol{\beta}}^{tr}(\lambda, r), \boldsymbol{A}^{tu}) \\
&= \frac{1}{n^{tu}} \sum_{j=1}^{p} I(\hat{\beta}^{tr}(\lambda, r)_j > 0) A^{tu}_{\cdot,j} - \frac{2}{n^{tu}} \sum_{i=1}^{n^{tu}} I\big(\sum_{j=1}^{p} A^{tu}_{ij} \hat{\beta}^{tr}(\lambda, r)_j > 0\big) \\
&\quad - \frac{\gamma}{\|\hat{\boldsymbol{\beta}}^{tr}(\lambda, r)\|_0} \sum_{j,k} c_{jk} I(\hat{\boldsymbol{\beta}}^{tr}(\lambda, r)_j > 0) I(\hat{\boldsymbol{\beta}}^{tr}(\lambda, r)_k > 0).
\end{aligned}$$

2.6.1. *Evaluation metrics.* Several metrics are used for evaluation, including the correct (C) or incorrect (IC) numbers of non-zero estimates for the mutations/genes in the true pathway B_0, and average differences of the cost function values (ADC) between the true set B_0 and the estimated set $\hat{B}$ of the driver mutations/genes; that is, C=$|B_0 \bigcap \hat{B}|$, IC=$|B_0^c \bigcap \hat{B}|$, ADC=$(f(B_0) - f(\hat{B}))/n$. We also included the running time (RT) (in minutes) of each algorithm. Note that ADC is important, because the basic task for minimum cost subset selection is to identify a set of mutations with the minimum cost.

In addition to using the correct (C) or incorrect (IC) numbers of non-zero estimates and ADC to measure how close the estimated pathways are close to the true pathway, we also investigate several other metrics in decomposing the cost function into the coverage (c_c) and exclusivity (c_e), and displaying the proportion of the patients carrying a mutation of a gene in a pathway (c_1), as well as the proportion of those carrying multiple mutations in more than one gene in the pathway (c_2). Specifically, we define

$$\begin{aligned}
f(B) &= c_e + c_c, \\
c_e &= \omega(B_0)/n, \qquad \hat{c}_e = \omega(\hat{B})/n, \\
c_c &= -|\Gamma(B_0)|/n, \qquad \hat{c}_c = -|\Gamma(\hat{B})|/n,
\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$c_1 = \sum_{i=1}^{n} I(\sum_{j\in B_0} A_{ij} = 1)/n, \qquad \hat{c}_1 = \sum_{i=1}^{n} I(\sum_{j\in \hat{B}} A_{ij} = 1)/n,$$
$$c_2 = \sum_{i=1}^{n} I(\sum_{j\in B_0} A_{ij} = 2)/n, \qquad \hat{c}_2 = \sum_{i=1}^{n} I(\sum_{j\in \hat{B}} A_{ij} = 2)/n.$$

Due to the coverage and exclusivity of a pathway, c_1 is often similar to $-c_c$ while c_2 is similar to c_e.

3. Results.

3.1. *Real data examples.* In this section we first illustrate the application of the proposed method to two cancer datasets that were previously examined by Vandin *et al.* (2012), then to a more recent and larger dataset including both mutation and expression data. As argued by Vandin *et al.* (2012), a set of mutated genes with a low cost function value is likely to be a mutated driver pathway, based on which our primary objective is to identify such mutated driver pathways through minimum cost subset selection of mutated genes. For each of the first two datasets, the proposed method was applied with the tuning parameter λ chosen from a tuning set of size 10, while 100 randomly generated initial estimates were used. For each initial estimate, we applied the proposed method, by which we identified multiple low-cost sets of mutations.

3.1.1. *Lung adenocarcinoma.* The original data set contains 1013 somatic mutations in 623 sequenced genes from 188 lung adenocarcinoma patients in the Tumor Sequencing Project (Ding *et al.*, 2008). For our purpose, we examined 356 genes that were mutated for at least one patient from a group of 162 patients, as in Vandin *et al.* (2012).

The proposed method was applied to identify multiple sets of mutated genes with low cost function values. Using 100 randomly selected initial values for MCSS, it cost 0.85 minutes and identified some gene sets with low cost. To demonstrate the resulting low-cost sets of mutations as possible candidates for mutated driver pathways, in Table 1 we group these discovered sets in terms of known pathways. In Table 1, all the discovered sets related to two known pathways associated with lung adenocarcinoma: the *mTOR* signaling pathway and the cell cycle pathway. Gene interactions in these pathways were reported in Ding *et al.* (2008) as depicted in Figure 1.

First, as indicated in Figure 1 (see Figure 6 of Ding *et al.* (2008)), the *mTOR* signaling pathway consists of some highly mutated genes, such as *EGFR, EPHA3, KRAS, NF1* and *STK11*. *EGFR* is a well-known oncogene,

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whose mutations are strongly associated with lung cancer (da Cunha Santos *et al.*, 2011). In contrast, *EPHA3* is one of the most frequently mutated genes in lung cancer, which however has not yet been extensively investigated. As suggested by Zhuang *et al.* (2012), tumor-suppressive effects of wild-type *EPHA3* could be overridden in trans by dominant negative *EPHA3* somatic mutations discovered in patients with lung cancer. *KRAS* is an oncogene associated with non-squamous non-small cell lung cancer. As indicated by many studies as well as our analysis, the mutations of *KRAS* and *EGFR* are strongly mutually exclusive. *KRAS* serves as a mediator between extracellular ligand binding and intracellular transduction of signals from the *EGFR* to the nucleus. The presence of activating KRAS mutations has been identified as a potent predictor of resistance to *EGFR*-directed antibodies (Heinemann *et al.*, 2009). *STK11* encodes a tumor suppressor enzyme, and its mutations can allow cells to grow and divide uncontrollably, leading to the formation of cancerous cells (Gill *et al.*, 2007). In particular, *STK11* mutations are found in non-squamous non-small cell lung cancer, however uncommon in most other types of cancer.

Interestingly, all the identified sets of mutated genes with the cost function values $f(\hat{B})$ lower than $-0.556 = 90/162$ are related to these five genes. Recall that in Ding *et al.* (2008), (*EGFR, KRAS*) ($f(\hat{B}) = -0.556$) and (*KRAS, STK11*) ($f(\hat{B}) = -0.420$) are the most significant pairs in the mutual exclusiveness test, and in Vandin *et al.* (2012), the triplet (*EGFR, KRAS, STK11*) ($f(\hat{B}) = -0.593$) was found with a lower cost, which was reported as a novel discovery. As indicated in Table 1, we could find not only this triplet (the second set in Table 1), but also another set (*EGFR, KRAS, NF1, STK11*) ($f(\hat{B}) = -0.611$) (the first set in Table 1) that contains this triplet and has a lower cost function value. It is a better characterized gene set, containing the already discovered (*EGFR, KRAS, STK11*). In addition, we also identified four low-cost sets: (*EGFR, KRAS, NF1*) ($f(\hat{B}) = -0.574$), (*EGFR, EPHA3, KRAS, NF1*) ($f(\hat{B}) = -0.574$), (*EGFR, EPHA3, KRAS*) ($f(\hat{B}) = -0.568$) and (*EGFR, KRAS*) ($f(\hat{B}) = -0.556$). These discoveries suggest possible roles of these genes related to the mTOR signaling pathway.

Second, the cell cycle pathway includes two highly mutated genes, *ATM* and *TP53*. *ATM* plays a central role in cell division and DNA repair, and the protein encoded by this gene is an important cell cycle checkpoint kinase, which functions as a regulator of a wide variety of downstream proteins. Some studies suggested that *ATM* mutations may increase the risk for lung cancer (Lo *et al.*, 2010). On the other hand, *TP53* encodes a tumor suppressor protein p53 that regulates cell division by keeping cells from growing and dividing too fast or in an uncontrolled way. *TP53* mutations are the

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 1. *Applied to the mutation data of lung adenocarcinoma (Ding et al., 2008), the new method MCSS identified multiple sets of low-cost mutated genes, grouped in terms of associated pathways.*

Pathway	Highly mutated genes	$\hat{B}$	$f(\hat{B})$	$\hat{c}_e$	$\hat{c}_c$	$\hat{c}_1$	$\hat{c}_2$
mTOR signaling	*EGFR, EPHA3, KRAS, NF1, STK11*	(*EGFR, KRAS, NF1, STK11*)	-0.611	0.117	-0.728	0.617	0.104
		(*EGFR, KRAS, STK11*)	-0.593	0.086	-0.679	0.593	0.086
		(*EGFR, KRAS, NF1*)	-0.574	0.031	-0.605	0.574	0.031
		(*EGFR, EPHA3, KRAS, NF1*)	-0.574	0.061	-0.636	0.586	0.037
		(*EGFR, EPHA3, KRAS*)	-0.568	0.025	-0.593	0.568	0.025
		(*EGFR, KRAS*)	-0.556	0	-0.556	0.556	0
cell cycle	*ATM, TP53*	(*ATM, TP53*)	-0.463	0.006	-0.469	0.463	0.006
mTOR signaling & cell cycle	*EGFR, EPHA3, KRAS, NF1, STK11* & ATM, TP53	(*ATM, EGFR, STK11,TP53*)	-0.525	0.173	-0.698	0.537	0.148
		(*KRAS, TP53*)	-0.469	0.148	-0.617	0.469	0.148
		(*EGFR, TP53*)	-0.444	0.068	-0.512	0.444	0.068

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

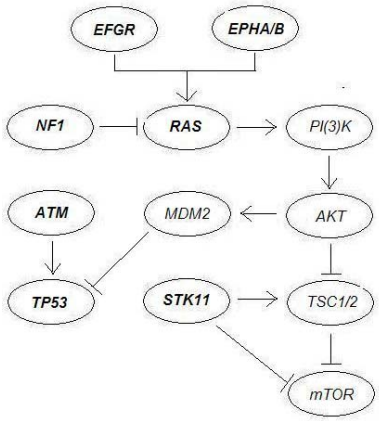

FIG 1. *The mTOR signaling pathway and the cell cycle pathway associated with lung adenocarcinoma as reported in Ding et al. (2008). The KRAS gene is one of the three oncogenes in the Ras family.*

most common genetic changes found in human cancer, in particular as one of the most significant events in lung cancer while playing an important role in the tumorigenesis of lung epithelial cells (Ding *et al.*, 2008).

The pair (*ATM, TP53*) was identified by the proposed method with the cost function value of -0.463, which was also discovered in Vandin *et al.* (2012) by removing the triplet (*EGFR, KRAS, STK11*) from the original dataset. Note that among the identified low-cost sets in Table 1, the cost function value of (*ATM, TP53*) was relatively high due to its low value of the coverage: $|\Gamma(\hat{B})| = 76$, much smaller than the maximum value of $n = 162$. As hypothesized in Vandin *et al.* (2012), the low coverage is possibly because somatic mutations were measured in only a small subset of genes, or because only single-nucleotide mutations and small indels in these genes were measured, and other types of genomic or epigenetic alterations might occur in the "unmutated" patients.

In addition, we identified some low-cost sets consisting of the genes related to both the mTOR signaling and the cell cycle pathways, namely, (*ATM, EGFR, STK11, TP53*) ($f(\hat{B}) = -0.525$), (*KRAS, TP53*) ($f(\hat{B}) = -0.469$) and (*EGFR, TP53*) ($f(\hat{B}) = -0.444$). Presumably these discoveries are related to that *EGFR* and *KRAS* are upstream regulators of *TP53*, as suggested by Ding *et al.* (2008).

3.1.2. *Glioblastoma multiforme (A).* Next, we analyzed the mutation data of 84 glioblastoma multiforme (GBM) patients from The Cancer

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome Atlas (The Cancer Genome Atlas Research Network, 2008), where 601 somatic mutations in these patients occurred. The mutation data consist of 84 patients and 178 genes, with each mutation occurring in at least one patient. The proposed method was applied to identify multiple sets of mutations with low cost values. Using 100 randomly selected initial values for MCSS, it cost 0.66 minutes and identified some gene sets with low cost. In Table 2 we also group the identified low-cost sets in terms of the possibly associated pathways. Most of the sets are associated with three important pathways of glioblastoma multiforme: the p53 signalling pathway, the RB signalling pathway and the RAS/RTK/PI(3)K signalling pathway. Interactions in these pathways were reported in The Cancer Genome Atlas Research Network (2008) as described in Figure 2. Below we discuss each pathway and the discovered sets of mutations.

First, the p53 signalling pathway consists of some highly mutated genes, *CDKN2A, MDM2, MDM4* and *TP53*. Importantly, mutations in the tumour suppressor gene *TP53* are typical events in primary glioblastoma multiforme, which is characterised by a short clinical history and the absence of a pre-existing, less malignant astrocytoma. In contrast, the cellular oncogene *MDM2* is viewed as an important negative regulator of the p53 tumor suppressor, whose overexpression is a characteristic feature of secondary glioblastoma multiforme, progressing from less malignant astrocytoma (Stark *et al.*, 2003).

Interestingly, the set of these four genes (*CDKN2A, MDM2, MDM4, TP53*) ($f(\hat{B}) = -0.655 = -55/84$) was identified by the proposed method as a novel discovery unreported before, e.g., in comparison with the pair (*CDKN2A, TP53*) ($f(\hat{B}) = -0.631$) identified by Vandin *et al.* (2012). As indicated in Table 2, the pair (*CDKN2A, TP53*) was also uncovered by the proposed method, in addition to another two sets, (*CDKN2A, DTX3, TP53*) ($f(\hat{B}) = -0.679$) and (*CDKN2B, TP53*) ($f(\hat{B}) = -0.631$). Since *CDKN2A* and *CDKN2B* are tumor suppressor genes located on a common homozygous deletion region on the human genome, they mutate almost simultaneously, which leads to a low cost function value of (*CDKN2B, TP53*). However, for (*CDKN2A, DTX3, TP53*), currently without further biological evidence, we conjecture that it has a low cost function value mainly because it consists of a low-cost set (*CDKN2A, TP53*) and gene *DTX3* with infrequent mutations.

Second, the RB signalling pathway consists of some highly mutated genes, *CDKN2A/B, CDK4, RB1*, where *CDKN2A* and *CDKN2B* are tumor suppressor genes, whose gene products, p16INK4A and p15INK4B, are both able to inhibit the binding of *CDK4* and *CDK6* to cyclin D, preventing

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the cell cycle progression at G1 phase. As a result, by negatively controlling cell cycle progression, these genes function as a critical defense against tumorigenesis of a great variety of human cancers, including glioblastoma multiforme (Feng *et al.*, 2012). The main set of mutations identified by the proposed method and associated with this pathway is likely to be (*CDKN2B, CYP27B1, RB1*) ($f(\hat{B}) = -0.738$) since it has very low cost and often overlaps with other sets with low cost, which is coincided with that identified by Vandin *et al.* (2012). Since the mutational profile of *CYP27B1* is nearly identical to a metagene including *CDK4*, Vandin *et al.* (2012) believed that the triplet (*CDKN2B, CDK4, RB1*) may be of interest. For (*CDKN2B, CYP27B1, RB1*), the low cost function value is mainly due to the inclusion of *CDKN2B* and *CYP27B1*. As shown in Table 2, we identified several other sets containing *CDKN2A*/*CDKN2B* and *CYP27B1*, namely, (*CDKN2A, CYP27B1, RB1*) ($f(\hat{B}) = -0.667$), (*CDKN2B, CYP27B1, NF1*) ($f(\hat{B}) = -0.667$), (*CDKN2A, CYP27B1, NF1*) ($f(\hat{B}) = -0.643$) and (*CDKN2B, CYP27B1*) ($f(\hat{B}) = -0.643$). In addition, we also uncovered a set (*CDKN2B, ERBB2, RB1, TSPAN31*) ($f(\hat{B}) = -0.762$), which is another new discovery by the proposed method. Interestingly, *TSPAN31* belongs to the same metagene including *CDK4*.

Third, the RAS/RTK/PI(3)K signalling pathway consists of some highly mutated genes, *EGFR, NF1, PI(3)K* and *PTEN*. Associated with this pathway, we identified a set of (*EGFR, KDR, NF1*) ($f(\hat{B}) = -0.619$). Its low cost function value is likely due to the inclusion of *EGFR* and *NF1*.

Finally, among the other identified low-cost sets in Table 2, (*MTAP, TP53, TSFM*) ($f(\hat{B}) = -0.667$), (*CYP27B1, MTAP, PTEN*) ($f(\hat{B}) = -0.655$) and (*CDK4, MTAP, PTEN*) ($f(\hat{B}) = -0.655$) are not known to be related to the pathways associated with glioblastoma multiforme. Hopefully, these low-cost sets will be useful for suggesting new links to glioblastoma multiforme. For (*EGFR, TP53*) ($f(\hat{B}) = -0.619$), its low cost function value is possibly due to the approximate exclusiveness of *EGFR* and *TP53*. In particular, tumors in the 'classical' subtype of glioblastoma multiforme often carry extra copies of *EGFR* and are rarely mutated in *TP53*.

In summary, as shown in the above two real data examples, nearly all of the identified low-cost sets by the proposed method are associated with some known mutated driver pathways. This suggests potential usefulness of the proposed method. More importantly, in comparison with an existing method, some new discoveries were obtained, such as (*EGFR, KRAS, NF1, STK11*) ($f(\hat{B}) = -0.611 = -99/162$) associated with the mTOR signalling pathway of lung cancer, and (*CDKN2A, MDM2, MDM4, TP53*) ($f(\hat{B}) = -0.656 = -55/84$) associated with the p53 signalling pathway of

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 2. *Applied to the mutation data of glioblastoma multiforme (data GBM A) (The Cancer Genome Atlas Research Network, 2008), the new method MCSS identified multiple sets of low-cost mutated genes, grouped in terms of associated pathways.*

Pathway	Highly mutated genes	$\hat{B}$	$f(\hat{B})$	$\hat{c}_e$	$\hat{c}_c$	$\hat{c}_1$	$\hat{c}_2$
p53 signalling	*CDKN2A, MDM2, MDM4, TP53*	(*CDKN2A, MDM2, MDM4, TP53*)	-0.655	0.167	-0.821	0.667	0.143
		(*CDKN2A, DTX3, TP53*)	-0.679	0.107	-0.786	0.691	0.083
		(*CDKN2A, TP53*)	-0.631	0.071	-0.702	0.631	0.071
		(*CDKN2B, TP53*)	-0.631	0.107	-0.738	0.631	0.107
RB signalling	*CDKN2A/B, CDK4, RB1*	(*CDKN2B, CYP27B1, RB1*)	-0.738	0.048	-0.786	0.738	0.048
		(*CDKN2B, ERBB2, RB1, TSPAN31*)	-0.762	0.071	-0.833	0.762	0.071
		(*CDKN2A, CYP27B1, RB1*)	-0.667	0.048	-0.714	0.667	0.048
		(*CDKN2B, CYP27B1, NF1*)	-0.667	0.107	-0.774	0.667	0.107
		(*CDKN2A, CYP27B1, NF1*)	-0.643	0.083	-0.723	0.643	0.083
		(*CDKN2B, CYP27B1*)	-0.643	0.036	-0.679	0.643	0.036
RAS signalling	*EGFR, NF1*	(*EGFR, KDR, NF1*)	-0.631	0.024	-0.655	0.631	0.024
Unknown		(*MTAP, TP53, TSFM*)	-0.667	0.131	-0.798	0.679	0.107
		(*CYP27B1, MTAP, PTEN*)	-0.655	0.155	-0.810	0.655	0.155
		(*CDK4, MTAP, PTEN*)	-0.655	0.155	-0.798	0.643	0.155
		(*EGFR, TP53*)	-0.619	0.083	-0.702	0.612	0.083

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

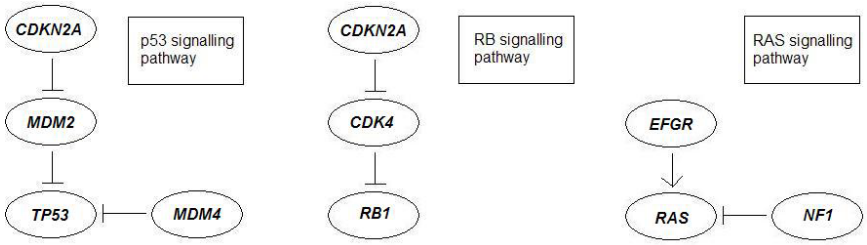

FIG 2. *Three pathways associated with glioblastoma multiforme as reported in The Cancer Genome Atlas Research Network (2008).*

glioblastoma multiforme.

3.1.3. *Glioblastoma multiforme (B).* Finally we analyzed a larger dataset of glioblastoma multiforme (GBM) patients from The Cancer Genome Atlas (Brennan *et al.*, 2013). The mutation data consist of 291 patients and 9539 genes, while the gene expression data include 558 patients and 12042 genes. Focusing on the intersection of the two gene sets, we obtained 5959 genes. Hence, we studied the filtered mutation data with 291 patients and 5959 genes, and the filtered expression data with 558 patients and 5959 genes.

First, the proposed MCSS was applied to identify multiple sets of mutations with low cost values using only the filtered mutation data. Using 10000 randomly selected initial values for all genes and 10000 randomly selected initial values for the subset of the genes with mutation rate larger than 0.05, MCSS identified some top gene sets with the six lowest cost function values (Table 3); note some gene sets with tied cost function values. They are mainly the variations and combinations of two core sets, (EGFR, KEL, NF1, TP53) and (IDH1, PIK3CA, PTEN), as contained in the top two sets identified. The list includes many well-known GBM genes, such as EGFR, PTEN, IDH1, TP53 and NF1 (Frattini *et al.*, 2013). Nevertheless, it is surprising that some top genes identified in Table 2 do not show up in the current list. Accordingly, we examined the top gene sets identified in the previous section but calculated their cost function values using the current data. From Table 4, we see that the top sets obtained earlier all have higher (i.e. worse) cost function values than those obtained in Table 3, indicating some inherent differences between the two datasets. For example, some high-mutation genes in the previous dataset, such as CDKN2A, MDM2, MDM4, CDKN2B, CYP27B1, ERBB2 and TSPAN31, had a low-mutation rate $<5\%$ in the current dataset. We use the less frequent mutation (LFM) (i.e. with a mutation rate $< 5\%$ among the subjects) ratio (i.e. the proportion of the

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 3

Application to the mutation data of glioblastoma multiforme (data GBM B) (Brennan et al., 2013): the top gene sets with the six lowest cost function values identified by the new method MCSS.

$\hat{B}$	$f(\hat{B})$	$\hat{c}_e$	$\hat{c}_c$	$\hat{c}_1$	$\hat{c}_2$
(*EGFR, KEL, NF1, CNTNAP2, TP53*)	-0.515	0.127	-0.642	0.526	0.106
(*EGFR, MUC4, KEL, CNTNAP2, TP53*)	-0.509	0.103	-0.612	0.512	0.096
(*FCGBP, IDH1, MUC16, PIK3CA, PTEN*)	-0.509	0.110	-0.619	0.512	0.103
(*EGFR, NF1, CNTNAP2, TP53*)	-0.505	0.103	-0.608	0.509	0.096
(*EGFR, MUC4, CNTNAP2, TP53, RYR3*)	-0.505	0.110	-0.615	0.509	0.103
(*FCGBP, IDH1, MUC16, PTEN*)	-0.498	0.065	-0.563	0.502	0.058
(*IDH1, MUC16, PIK3CA, PTEN*)	-0.498	0.089	-0.587	0.498	0.089
(*DSP, MUC4, FCGBP, IDH1, NF1, MUC16, PTEN*)	-0.498	0.175	-0.673	0.512	0.148
(*IDH1, NF1, MUC16, PTEN*)	-0.495	0.096	-0.591	0.495	0.096
(*EGFR, KEL, TP53, FLG*)	-0.491	0.131	-0.622	0.502	0.110
(*EGFR, MUC4, CNTNAP2, TP53*)	-0.491	0.083	-0.574	0.491	0.082
(*EGFR, USH2A, CNTNAP2, TP53*)	-0.491	0.096	-0.587	0.498	0.082
(*DSP, IDH1, MUC16, DNAH3, PTEN*)	-0.491	0.100	-0.591	0.495	0.093
(*ATRX, FCGBP, MUC16, PIK3CA, PTEN*)	-0.491	0.124	-0.615	0.502	0.103
(*EGFR, CNTNAP2, TP53, RYR3*)	-0.491	0.089	-0.581	0.491	0.089
(*EGFR, IDH1, NF1, MUC16, RELN*)	-0.491	0.110	-0.601	0.495	0.103

LFM genes in a gene set) to indicate the presence of LFM genes in Table 4. The inherent differences between the two datasets confirm the genomic heterogeneity of GBM, one of the biggest challenges in current data analysis.

Finally, MCSS_ME was applied in an integrative analysis of both the filtered mutation and gene expression data. We did not apply the GA method because its current implementation requires the same set of the subjects with both mutation and gene expression data, which did not hold here. Using 10000 randomly selected initial values, MCSS_ME identified its top 10 gene sets shown in Table 5. We note that several genes were also identified

TABLE 4

The cost function values of the gene sets in the larger GBM (B) dataset with the gene sets identified from the smaller GBM (A) dataset.

$\hat{B}$	$f(\hat{B})$	$\hat{c}_e$	$\hat{c}_c$	$\hat{c}_1$	$\hat{c}_2$	LFM ratio
(*CDKN2A, MDM2, MDM4, TP53*)	-0.285	0.007	-0.292	0.285	0.007	3/4
(*CDKN2A, TP53*)	-0.289	0	-0.289	0.289	0	1/2
(*CDKN2B, TP53*)	-0.285	0	-0.285	0285	0	1/2
(*CDKN2B, CYP27B1, RB1*)	-0.103	0	-0.103	0.103	0	2/3
(*CDKN2B, ERBB2, RB1, TSPAN31*)	-0.103	0	-0.103	0.103	0	3/4
(*CDKN2A, CYP27B1, RB1*)	-0.107	0	-0.107	0.107	0	2/3
(*CDKN2B, CYP27B1, NF1*)	-0.124	0	-0.124	0.124	0	2/3
(*CDKN2A, CYP27B1, NF1*)	-0.124	0	-0.124	0.124	0	2/3
(*CDKN2B, CYP27B1*)	-0.127	0	-0.127	0.127	0	2/2
(*EGFR, KDR, NF1*)	-0.354	0.024	-0.378	0.354	0.024	1/3
(*EGFR, TP53*)	-0.447	0.048	-0.495	0.447	0.048	0/2

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the other dataset in the previous section. Many selected genes are annotated in the Cancer Gene Census in the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes *et al.*, 2015), including well-known GBM genes (EGFR, PTEN, IDH1, TP53 and NF1, among others) (Frattini *et al.*, 2013). Here we only highlight a few examples. Gene ATRX was an important member of the H3.3-ATRX-DAXX chromatin remodelling pathway, among the most frequently mutated genes in paediatric and adult GBM Schwartzentruber *et al.* (2012). Gene PIK3CA, encoding a protein that antagonizes the function of PTEN protein in the PI3K/Akt pathway; an exclusive mutation pattern was observed in PIK3CA and PTEN (Hartmann *et al.* , 2005). Mutations in a single gene, IDH1, resulted in reorganization of the methylome and transcriptome in glioblastomas and other cancers (Turcan *et al.*, 2012). As reviewed in Sturm *et al.* (2014), unsupervised clustering of the gene expression data from 200 adult GBM samples from TCGA identified four different molecular subtypes: proneural, neural, classical and mesenchymal. The proneural subtype was largely characterized by abnormalities in platelet derived growth factor receptor α (PDGFRA) or isocitrate dehydrogenase 1 (IDH1), whereas mutation of the epidermal growth factor receptor (EGFR) was found in the classical subgroup and mutations in neurofibromin (NF1) were common in mesenchymal tumours. In particular, Sturm *et al.* (2014) mentioned the detection of lower-frequency events in both cancer-related as well as previously un-associated genes such as ATRX and KEL.

Note that all the gene sets identified with only the mutation data include only high-mutation genes (i.e. with a mutation rate $> 5\%$ among the subjects), while it is of interest but difficult to identify driver genes with less frequent mutations (i.e. with a mutation rate $\leq 5\%$) (LFM). Hence, we show the LFM ratio in Table 5. It is interesting to note the presence of two LFM genes, CNTP2 and DH3. In summary, our preliminary results seem to support the use of integrative analysis as advocated by others (Frattini *et al.*, 2013).

3.2. *Simulations.* Due to the difficulties in evaluating de novo discoveries with real data, we performed extensive simulations to study the operating characteristics of the proposed method and compared its performance against its competitors. All simulations were performed on a single processor of an Intel(R) Xeon(R) 2.83GHz PC.

3.2.1. *Simulation I: a single driver pathway.* We first considered the case with only a single driver pathway, in which the focus was on comparing our new method with its strong competitor, the MCMC algorithm of Dendrix as implemented in `Python` (Vandin *et al.*, 2012), though several other methods

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 5. *Application to the mutation data and gene expression data of glioblastoma multiforme (Brennan et al., 2013): the top 10 gene sets identified by the new method MCSS_ME with the automatically selected $\gamma = 0.1$. The known cancer genes annotated on COSMIC are underlined.*

$\hat{B}$	$f_{ME}(\hat{B})$	$\hat{c}_e$	$\hat{c}_c$	$\hat{c}_1$	$\hat{c}_2$	$f(\hat{B})$	$\gamma f_E(\hat{B})$	LFM ratio
(*FCGBP*, *RYR2*, *PCLO*, *CNTP2*, *TP53*)	-0.781	0.113	-0.515	0.419	0.079	-0.402	-0.379	1/5
(*ATRX*, *PIK3CA*, *DOCK5*, *MUC5B*, *DH3*, *PTEN*)	-0.779	0.113	-0.519	0.419	0.086	-0.405	-0.374	1/6
(*EGFR*, *KEL*, *NF1*, *TP53*, *DH3*)	-0.761	0.144	-0.625	0.498	0.113	-0.481	-0.280	0/5
(*PIK3CA*, *TP53*, *PTEN*)	-0.754	0.137	-0.549	0.419	0.124	-0.412	-0.342	0/3
(*PIK3R1*, *DSP*, *MUC4*, *FCGBP*, *MUC16*, *PTEN*)	-0.753	0.162	-0.612	0.474	0.113	-0.450	-0.303	0/6
(*ATRX*, *KEL*, *PIK3CA*, *PTEN*)	-0.752	0.052	-0.474	0.423	0.052	-0.423	-0.329	0/4
(*DSP*, *FCGBP*, *IDH1*, *MUC16*, *DOCK5*, *PTEN*)	-0.751	0.124	-0.612	0.502	0.096	-0.488	-0.263	0/6
(*KEL*, *PIK3CA*, *FRAS1*, *MUC5B*, *DH3*, *PTEN*)	-0.745	0.117	-0.529	0.426	0.089	-0.413	-0.332	1/6
(*FCGBP*, *IDH1*, *MUC16*, *PIK3CA*, *PTEN*)	-0.740	0.110	-0.619	0.512	0.103	-0.509	-0.231	0/5
(*EGFR*, *IDH1*, *KEL*, *NF1*, *PIK3CA*, *DNAH3*, *PTEN*)	-0.739	0.244	-0.701	0.495	0.168	-0.457	-0.282	0/7

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were also included.

For the proposed method, we fixed $\tau_1 = 1$, $\tau_2 = 0.1$ and $\alpha = 10^{-3}$, and tuned λ over a tuning set Λ. Specifically, λ was selected by minimizing a tuning error over a set of 10 equally-spaced points. We used 100 random initial estimates for MCSS (based on the subgradient descent algorithm), containing the Lasso estimate $\hat{\boldsymbol{\beta}}_L$, as well as the other 99 random initial estimates. For the algorithm of Dendrix Vandin *et al.* (2012), 1000000 iterations were run for MCMC with sampling sets of size 4 for every 1000 iterations. Moreover, the algorithm was run with the number of driver mutations varying from 1 to 10 to select the best fitted subset with the lowest cost of $f(\cdot)$ in (2.1) as the final result.

For each simulated dataset, an $n \times p$ mutation matrix $\boldsymbol{A}$ was generated with a 1 indicating a mutation and 0 otherwise. For each patient, a gene in a driver pathway $B_0 = \{1, 2, 3, 4\}$ was randomly selected and it mutated with probability p_1, and another gene in B_0 was randomly selected to have a mutation with probability p_2. Consequently, p_1 and p_2 controlled the coverage and exclusiveness of B_0 respectively. Other genes outside B_0 mutated with probability p_3. Six set-ups were examined with $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$: (1) $n = 50$ and $p = 1000$, (2) $n = 100$ and $p = 1000$, (3) $n = 1000$ and $p = 50$, (4) $n = 1000$ and $p = 100$, (5) $n = 50$ and $p = 10000$, (6) $n = 100$ and $p = 10000$. With $(p_1, p_2, p_3) = (0.8, 0.02, 0.05)$, we had similar set-ups. The simulation results are summarized in Tables 6 and 7.

As suggested in Tables 6 and 7, the proposed method outperformed the MCMC algorithm of Dendrix, especially in the high-dimensional situations, with respect to the accuracy of selection as well as computational efficiency measured by the values of C, IC, ADC and RT respectively. The amount of improvement of the proposed method over the competitor ranged from low to high. For the running time, the proposed algorithm was overwhelmingly faster than the MCMC algorithm of Dendrix. In particular, it was often more than 50 times faster than the MCMC algorithm of Dendrix. As expected, both methods tended to perform worse as the amount of coverage and exclusiveness of a mutated driver pathway decreased.

We also compared our new method with several other alternative methods that were proposed more recently, including Multi-dendrix-MCMC of Leiserson *et al.* (2013), BLP (binary linear programming) and GA (genetic algorithm) of Zhao *et al.* (2012). The numerical results of the three methods are also summarized in Tables 6 and 7. These results suggest that the performance of Multi-dendrix-MCMC was quite similar to that of Dendrix-MCMC but much faster; BLP and GA performed better than their competitors if the algorithms could finish running; however, they were not robust with frequent

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 6. *Results in Simulation I based on 100 simulation replications with* $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$. *The sample means (SD in parentheses) of correct (C) or incorrect (IC) numbers of non-zero estimates, average differences of the cost (ADC) between the true gene subset* $B_0 = \{1, 2, 3, 4\}$ *and the estimated subset* $\hat{B}$, *that is,* $\frac{f(B_0)-f(\hat{B})}{n}$, *and the running time (RT) (in minutes) of the algorithms.*

n	p	Method	C	IC	ADC	$\hat{c}_1$ [c_1]	$\hat{c}_2$ [c_2]	RT
50	1000	MCSS	4 (0)	0 (0)	0 (0)	.95 [.95]	.01 [.00]	.22 (.02)
		Dendrix-MCMC	3.80 (.41)	.50 (.94)	-.02 (.04)	.94 [.95]	.01 [.00]	16.89 (2.01)
		Multi-dendrix-MCMC	3.90 (.30)	.15 (.36)	-.01 (.03)	.95 [.95]	.01 [.00]	.81 (.01)
		BLP	3.39 (.86)	3.82 (2.59)	.05 (.03)	.99 [.95]	.00 [.00]	.01 (.01)
		GA	3.90 (.38)	2.13 (1.69)	.04 (.03)	.98 [.95]	.01 [.00]	2.97 (.25)
100	1000	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.37 (.04)
		Dendrix-MCMC	4 (0)	1.00 (.53)	.02 (.01)	.95 [.94]	.01 [.01]	27.01 (3.71)
		Multi-dendrix-MCMC	4 (0)	1.1 (.55)	.03 (.02)	.98 [.94]	.01 [.01]	.81 (.01)
		BLP	4 (0)	1.76 (1.27)	.01 (.01)	.97 [.94]	.02 [.01]	.05 (.01)
		GA	4 (0)	1.68 (1.21)	.01 (.01)	.97 [.94]	.02 [.01]	1.96 (.16)
1000	50	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.27 (.01)
		Dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	134.34 (27.79)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.78 (.19)
		BLP	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.07 (.01)
		GA	0 (0)	0 (0)	-.94 (.00)	0 [.94]	0 [.01]	.00 (.00)
1000	100	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.41 (.03)
		Dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	144.46 (28.75)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.68 (.24)
		BLP	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.21 (.01)
		GA	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.11 (.00)
50	10000	MCSS	4 (0)	0 (0)	0 (0)	.95 [.95]	.01 [.01]	1.67 (.29)
		Dendrix-MCMC	1.25 (1.02)	5.96 (3.62)	-.24 (.04)	.83 [.95]	.02 [.01]	67.06 (5.22)
		Multi-dendrix-MCMC	1.45 (1.23)	5.25 (4.02)	-.24 (.03)	.83 [.95]	.03 [.00]	1.88 (.03)
		BLP	3.42 (.91)	2.72 (2.06)	.05 (.03)	.99 [.95]	.01 [.00]	.27 (.48)
		GA	3.93 (.25)	1.50 (.92)	.04 (.02)	.98 [.95]	.01 [.00]	284.92 (32.79)
100	10000	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	3.94 (.43)
		Dendrix-MCMC	4 (0)	.96 (.41)	.02 (.01)	.95 [.94]	.01 [.01]	75.46 (8.37)
		Multi-dendrix-MCMC	4 (0)	.90 (.44)	.02 (.01)	.96 [.94]	.01 [.01]	3.75 (.02)
		BLP	3.95 (.22)	3.90 (1.58)	.04 (.02)	.99 [.94]	.01 [.01]	2.89 (2.96)
		GA	3.97 (.14)	3.45 (1.51)	.04 (.02)	.99 [.94]	.01 [.01]	275.39 (69.92)

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 7. *Results in Simulation I based on* 100 *simulation replications with* $(p_1, p_2, p_3) = (0.8, 0.02, 0.05)$.

n	p	Method	C	IC	ADC	$\hat{c}_1$ [c_1]	$\hat{c}_2$ [c_2]	RT
50	1000	MCSS	3.70 (.47)	.25 (.44)	-.01 (.02)	.79 [.79]	.02 [.01]	.17 (.02)
		Dendrix-MCMC	2.95 (.83)	3.15 (2.21)	-.03 (.04)	.77 [.79]	.01 [.01]	16.39 (2.22)
		Multi-dendrix-MCMC	3.00 (.72)	3.40 (1.93)	-.01 (.02)	.79 [.79]	.01 [.01]	.59 (.01)
		BLP	2.85 (1.01)	6.62 (1.53)	.20 (.05)	1.00 [.79]	.00 [.01]	.04 (.02)
		GA	3.63 (.61)	5.21 (1.04)	.18 (.04)	.97 [.79]	.02 [.01]	2.58 (.21)
100	1000	MCSS	4 (0)	.05 (.07)	.00 (.00)	.79 [.79]	.01 [.01]	.30 (.05)
		Dendrix-MCMC	4 (0)	2.40 (.60)	.05 (.05)	.84 [.79]	.01 [.01]	29.60 (1.72)
		Multi-dendrix-MCMC	4 (0)	2.25 (.78)	.05 (.01)	.85 [.79]	.05 [.01]	.82 (.02)
		BLP	3.89 (.31)	5.90 (.70)	.11 (.02)	.91 [.79]	.04 [.01]	.09 (.01)
		GA	4 (0)	5.65 (.67)	.11 (.03)	.90 [.79]	.04 [.01]	2.00 (.11)
1000	50	MCSS	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	.27 (.02)
		Dendrix-MCMC	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	163.85 (30.12)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	1.81 (.34)
		BLP	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	.09 (.01)
		GA	0 (0)	0 (0)	-.78 (.01)	0 [.78]	0 [.02]	.00 (.00)
1000	100	MCSS	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	.36 (.03)
		Dendrix-MCMC	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	186.10 (37.15)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	1.51 (.50)
		BLP	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	.27 (.03)
		GA	4 (0)	0 (0)	0 (0)	.78 [.78]	.02 [.02]	.11 (.00)
50	10000	MCSS	3.15 (.67)	.40 (.68)	-.06 (.05)	.73 [.79]	.02 [.01]	.98 (.27)
		Dendrix-MCMC	.30 (.42)	8.70 (1.16)	-.10 (.05)	.71 [.79]	.03 [.01]	56.09 (4.85)
		Multi-dendrix-MCMC	.15 (.48)	9.10 (1.07)	-.05 (.02)	.77 [.79]	.03 [.01]	1.86 (.03)
		BLP	2.45 (1.40)	5.44 (2.13)	.20 (.05)	.99 [.79]	.00 [.01]	.83 (1.06)
		GA	3.24 (1.20)	3.82 (2.16)	.16 (.04)	.95 [.79]	.01 [.01]	276.33 (37.50)
100	10000	MCSS	4 (0)	.05 (.07)	.00 (.00)	.79 [.79]	.01 [.01]	2.70 (.46)
		Dendrix-MCMC	3.98 (.00)	1.87 (.54)	.01 (.01)	.79 [.79]	.01 [.01]	64.89 (6.75)
		Multi-dendrix-MCMC	3.64 (.99)	1.17 (2.35)	.12 (.28)	.94 [.79]	.01 [.01]	3.82 (.04)
		BLP	3.85 (.36)	6.05 (.51)	.17 (.02)	.98 [.79]	.00 [.01]	12.43 (8.76)
		GA	4 (0)	5.33 (.62)	.15 (.02)	.96 [.79]	.01 [.01]	252.74 (73.58)

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

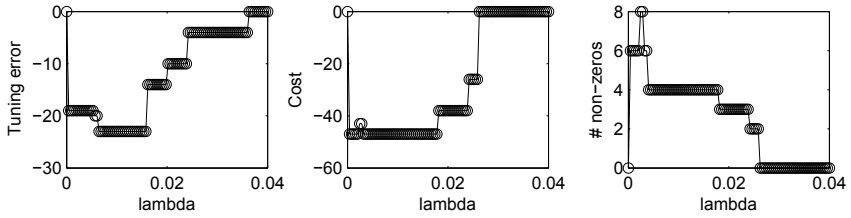

FIG 3. *Tuning error, cost and number of non-zero (i.e. true positive) estimates of MCSS versus the tuning parameter value* $\lambda \in \Lambda$ *with* $|\Lambda| = 100$ *for the first simulation set-up:* $n = 50$, $p = 1000$ *and* $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$.

running errors (up to 15% failing to converge or giving output properly). In particular, BLP ran quite unsteadily in high-dimensional situations, say $n = 50$ and $p = 1000$ or 10000, while GA was too slow in high-dimensional situations since it tried to seek an exact solution. As expected, we see that these three methods also tended to perform worse as the amount of coverage and exclusiveness of a mutated driver pathway decreased. Since a rarely mutated gene may by chance satisfy the (approximate) exclusivity property with a highly mutated gene, the union of the highly mutated gene and some rarely mutated genes could drive down the cost function value, leading to false positives. To investigate this issue, we conducted a simulation study. As before, the driver pathway contained four genes. We set the 1st gene to have mutation in a fraction p_0^* of all n patients, while setting the other three driver genes $\{2, 3, 4\}$ to have mutations only in the remaining patients, for whom a gene from $\{2, 3, 4\}$ was randomly selected with probability p_1^* to have a mutation, and another gene in $\{2, 3, 4\}$ was randomly selected to have a mutation with probability p_2^*. Finally, other genes outside $B_0 = \{1, 2, 3, 4\}$ mutated with a background probability p_3^*. The corresponding simulation results are summarized in Table 8, suggesting that the proposed method still performed well.

To evaluate the performance involving cross-validation, consider the first set-up: $n = 50$ and $p = 1000$ with $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$. The cross-validation procedure was applied with an enlarged size of Λ, say 100, and Algorithm 1 was applied to A for each $\lambda \in \Lambda$ separately. The results are displayed in Figure 3, demonstrating that the λ's minimizing the tuning error corresponded to the minimum cost of (2.1) and the true size of B_0, say 4.

Moreover, the current tuning error is obtained by applying the cross-validation procedure for once in consideration of computational efficiency.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 8. *Results in Simulation I based on 100 simulation replicates with for* $(p_0^*, p_1^*, p_2^*, p_3^*) = (0.7, 0.8, 0.02, 0.05)$.

n	p	Method	C	IC	ADC	$\hat{c}_1[c_1]$	$\hat{c}_2[c_2]$	RT
50	1000	MCSS	2.95 (.60)	.20 (.52)	-.05 (.05)	.88 [.94]	.02 [.00]	.48 (.05)
		Dendrix-MCMC	2.60 (.50)	1.55 (1.27)	-.04 (.03)	.90 [.94]	.01 [.00]	12.58 (.78)
		Multi-dendrix-MCMC	2.45 (.75)	2.20 (1.91)	-.04 (.03)	.90 [.94]	.02 [.00]	.46 (.05)
		BLP	3.32 (.81)	2.95 (1.50)	.06 (.03)	1.00 [.94]	.00 [.00]	.02 (.01)
		GA	3.15 (.93)	3.25 (1.58)	.06 (.03)	1.00 [.94]	.00 [.00]	2.50 (.18)
100	1000	MCSS	3.85 (.41)	.30 (.73)	-.02 (.05)	.91 [.94]	.03 [.00]	.64 (.08)
		Dendrix-MCMC	3.95 (.34)	.31 (.61)	-.00 (.01)	.94 [.94]	.01 [.00]	21.55 (0.84)
		Multi-dendrix-MCMC	3.95 (.22)	.35 (.93)	-.00 (.01)	.94 [.94]	.01 [.00]	.70 (.07)
		BLP	3.95 (.22)	3.00 (1.72)	.03 (.02)	.97 [.94]	.01 [.00]	.05 (.02)
		GA	4 (0)	2.60 (1.39)	.03 (.01)	.97 [.94]	.01 [.00]	1.84 (.11)
1000	50	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.49 (.04)
		Dendrix-MCMC	4 (0)	0.10 (.31)	-.00 (.01)	.93 [.94]	.01 [.01]	127.60 (5.49)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.43 (.21)
		BLP	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.04 (.00)
		GA	0 (0)	0 (0)	-.94 (.01)	0 [.94]	0 [.01]	.00 (.00)
1000	100	MCSS	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.65 (.06)
		Dendrix-MCMC	4 (0)	.05 (.22)	-.00 (.01)	.94 [.94]	.01 [.01]	153.12 (7.03)
		Multi-dendrix-MCMC	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	1.16 (.42)
		BLP	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.08 (.00)
		GA	4 (0)	0 (0)	0 (0)	.94 [.94]	.01 [.01]	.10 (.01)
50	10000	MCSS	2.65 (.68)	.95 (1.10)	-.10 (.05)	.86 [.94]	.04 [.00]	5.91 (1.52)
		Dendrix-MCMC	1.25 (.44)	3.20 (1.32)	-.07 (.03)	.87 [.94]	.01 [.00]	40.77 (2.01)
		Multi-dendrix-MCMC	1.50 (.57)	3.75 (2.51)	-.08 (.04)	.87 [.94]	.02 [.01]	1.37 (.09)
		BLP	2.89 (1.17)	2.40 (1.27)	.06 (.03)	1.00 [.94]	.00 [.00]	.15 (.31)
		GA	2.00 (1.08)	3.70 (1.62)	.06 (.03)	1.00 [.94]	.00 [.00]	224.29 (30.99)
100	10000	MCSS	3.15 (.64)	0 (0)	-.04 (.03)	.89 [.94]	.00 [.00]	10.92 (2.57)
		Dendrix-MCMC	2.70 (.57)	1.41 (2.06)	-.06 (.02)	.88 [.94]	.02 [.00]	49.40 (1.76)
		Multi-dendrix-MCMC	3.34 (.57)	2.04 (2.65)	-.04 (.02)	.90 [.94]	.02 [.00]	2.41 (.15)
		BLP	3.30 (.92)	5.45 (1.90)	.06 (.02)	1.00 [.94]	.00 [.00]	.42 (.24)
		GA	3.65 (.67)	4.50 (1.50)	.06 (.02)	.99 [.94]	.00 [.00]	308.46 (44.10)

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

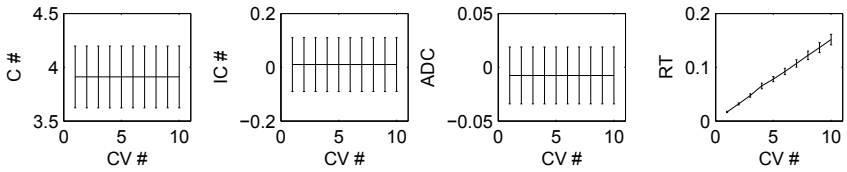

FIG 4. *The correct (C #) and incorrect (IC #) numbers of non-zero (i.e. true positive) estimates, the average difference of the costs (ADC) and running time (RT) of MCSS versus the fold number of cross-validation used in Algorithm 2 (CV #) for the first simulation set-up:* $n = 50$, $p = 1000$ *and* $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$.

TABLE 9
Results in Simulation II based on 100 *simulation replications with* $(p_1, p_2, p_3) = (0.8, 0.02, 0.05)$.

n	p	Method	C	IC	ADC	RT
50	1000	Multi-dendrix-MCMC	2.70 (1.65)	13.40 (6.09)	-.27 (.07)	.59 (.02)
		MCSS	7.35 (.67)	.55 (.89)	-.03 (.06)	.35 (.06)
100	1000	Multi-dendrix-MCMC	8 (0)	2.80 (1.19)	.05 (.01)	.85 (.01)
		MCSS	8 (0)	0 (0)	0 (0)	.57 (.06)
50	10000	Multi-dendrix-MCMC	.25 (.55)	18.55 (1.73)	-.37 (.08)	1.88 (.05)
		MCSS	5.52 (1.54)	1.91 (1.89)	-.14 (.10)	3.69 (.38)
100	10000	Multi-dendrix-MCMC	5.75 (1.06)	2.75 (3.91)	-.25 (.07)	3.87 (.04)
		MCSS	7.42 (.82)	.65 (.67)	-.14 (.13)	7.15 (1.11)

For instance, in the first set-up with $(n = 50, p = 1000)$ and $(p_1, p_2, p_3) = (0.95, 0.01, 0.05)$, as indicated in Figure 4, as the cross-validation fold number increased, the performance of the proposed method measured by C, IC and ADC did not improve, while RT increased linearly.

3.2.2. *Simulation II: multiple driver pathways.* We further compared the performance of MCSS against Multi-dendrix in identifying multiple true driver pathways as follows.

The simulation set-up was similar as before except that there were two true driver pathways B_1 and B_2. We used 100 random initial estimates for MCSS. We compared their performance using the top two estimated sets (with the minimum cost function values) by each method for each dataset. As shown in Table 9, MCSS performed much better for the most challenging high-dimensional case with $p = 10000$ and $n = 50$: it correctly identified a much larger number of the genes in the two true driver pathways (i.e. with a larger number of estimated true positives) while yielding fewer false positives. On the other hand, as the sample size n increased to 100, the performance of Multi-dendrix caught up.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2.3. *Simulation III: with both mutation and gene expression data.* We generated the mutation data as in Table 7 and the gene expression data from a multivariate normal distribution $N(0, V)$. Specifically, we divided the genes $\{1, \cdots, p\}$ into mutually disjoint subsets $B_0 = \{1, 2, 3, 4\}$, B_1, B_2, ...B_K, where for each $k \in \{1, \cdots, K\}$, the gene set size $|B_k|$ was random from $\{2, \cdots, 20\}$. V is a correlation matrix with all diagonal elements $V_{jj} = 1$; for any $j_1 < j_2$ both in the same B_k, $V_{j_1 j_2} = V_{j_2 j_1} = 0.9$; otherwise, $V_{j_1 j_2} = V_{j_2 j_1} = 0.1$. The rationale is that, for the genes in the same set, due to their shared function, their expression levels are also highly correlated. We used our proposed method to select all the tuning parameters, including γ. The simulation results for the integrative analysis of both mutation data and gene expression data are summarized in Table 10, where the new method MCSS_ME is compared with GA_ME, the integrative version of GA (Zhao *et al.*, 2012). Note that, to our knowledge, the integrative version of BLP in Zhao *et al.* (2012) is not yet publicly available. From Table 10, we see that GA_ME failed in situations with the dimension p much smaller than the sample size n; in contrast, the new method MCSS_ME performed well. Furthermore, GA_ME was much time-consuming for large p.

4. Conclusions. This paper has introduced a new computational method for a combinatorial optimization problem motivated from cancer genomics. It approximates a combinatorial cost function with a continuous and non-convex relaxation. In particular, the indicator function is approximated by a non-convex truncated L_1-function. The proposed method is computationally more efficient than an existing approach based on stochastic search, and compares favorably over several existing methods in simulations. Through both real data and simulated data analyses, the proposed method was shown to be promising for discovering mutated driver pathways with tumor sequencing data. In light of that Dendrix and other methods have been successfully applied to the TCGA (Kandoth *et al.*, 2013), it would be interesting to apply our proposed method to on-going large cancer genomics projects. Furthermore, the current problem differs from existing pathway analysis of genome-wide association studies (GWAS) (Wang *et al.*, 2007; Torkamani *et al.*, 2007; Schaid *et al.*, 2012) in two aspects: (i) the current problem is more challenging in the sense that no pathway is given a priori; (ii) however, GWAS data is different with genetic variants (or mutations) present for healthy control subjects, and it is also higher-dimensional with a larger number of genetic variants. It would be interesting to see whether the key concept of mutation exclusivity and associated methodology in the current context can be extended and applied to GWAS for de novo pathway

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE 10. *Results in Simulation III for integrative analysis of mutation data and gene expression data.* $(p_1, p_2, p_3) = (0.8, 0.02, 0.05)$

n	p	Method	C	IC	ADC	RT
50	1000	MCSS_ME	4 (0)	0 (0)	0 (0)	5.11 (.38)
		GA_ME	3.61 (.54)	5.62 (1.67)	-.36 (.04)	38.22 (1.95)
100	1000	MCSS_ME	4 (0)	0 (0)	0 (0)	7.96 (.70)
		GA_ME	4 (0)	5.6 (.54)	-.38 (.05)	30.81 (1.28)
1000	50	MCSS_ME	4 (0)	0 (0)	0 (0)	.58 (.02)
		GA_ME	0 (0)	0 (0)	-1.41 (.01)	.00 (.00)
1000	100	MCSS_ME	4 (0)	0 (0)	0 (0)	.77 (.03)
		GA_ME	4 (0)	0 (0)	0 (0)	2.10 (.12)
50	10000	MCSS_ME	4 (0)	0 (0)	0 (0)	432.55 (35.63)
		GA_ME	– (–)	– (–)	– (–)	> 1500.00 (–)
100	10000	MCSS_ME	4 (0)	0 (0)	0 (0)	100.77 (24.25)
		GA_ME	– (–)	– (–)	– (–)	> 1500.00 (–)

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or gene subnetwork (Liu *et al.*, 2014) discovery to handle genetic heterogeneity. Finally, the main idea of our algorithm is quite general and may be modified and extended for other challenging combinatorial search problems.

Matlab code implementing the new method and a manual are available at `https://github.com/ChongWu-Biostat/MCSS` .

APPENDIX

Proof of Theorem 1. For convergence of Algorithm 1, by construction, we have, for $m \in \mathbb{N}$, $S(\hat{\boldsymbol{\beta}}^{(m)}) = S^{(m+1)}(\hat{\boldsymbol{\beta}}^{(m)}) \leq S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)}) \leq S^{(m)}(\hat{\boldsymbol{\beta}}^{(m-1)}) = S(\hat{\boldsymbol{\beta}}^{(m-1)})$. Since $S(\boldsymbol{\beta})$ is obviously bounded below, the convergence is proved. Converging finitely follows from the strict decreasing character of $S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)})$ in m, uniqueness of minimizer of $S^{(m)}(\boldsymbol{\beta})$ and finite possible values of $\nabla S_2(\hat{\boldsymbol{\beta}}^{(m-1)})$ in (2.4). After termination occurs at $m^\star$, $\hat{\boldsymbol{\beta}}^{(m)}$ remains unchanged for $m \geq m^\star$, so does the cost function $S(\hat{\boldsymbol{\beta}}^{(m)})$ in (2.3) for $m \geq m^\star$. By construction of $S(\boldsymbol{\beta})$, we have that $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m^\star-1)}$, for all $m \geq m^\star$. $\tilde{\boldsymbol{\beta}}$ is uniquely defined, because for each $m \in \mathbb{N}$, the minimizer $\hat{\boldsymbol{\beta}}^{(m)}$ of $S^{(m)}(\beta)$ is uniquely defined. Since $\nabla S^{(m^\star)}(\hat{\boldsymbol{\beta}}^{(m^\star)}) = \nabla S_1(\hat{\boldsymbol{\beta}}^{(m^\star)}) - \nabla S_2(\hat{\boldsymbol{\beta}}^{(m^\star-1)}) = 0$, we get that $\nabla S_1(\hat{\boldsymbol{\beta}}^{(m^\star)}) = \nabla S_2(\hat{\boldsymbol{\beta}}^{(m^\star-1)}) = \nabla S_2(\hat{\boldsymbol{\beta}}^{(m^\star)})$. Thus, $\nabla S(\hat{\boldsymbol{\beta}}^{(m^\star)}) = \nabla S_1(\hat{\boldsymbol{\beta}}^{(m^\star)}) - \nabla S_2(\hat{\boldsymbol{\beta}}^{(m^\star)}) = 0$, which completes the proof.

Proof of Lemma 1. We prove by contradiction. By construction of $S(\boldsymbol{\beta})$, we see that $|\boldsymbol{\beta}^*| = (|\beta_1^*|, \cdots, |\beta_p^*|)^T$ is also a local minimum of $S(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in R^p$. Without loss of generality, we assume that $|\beta_1^*| > \tau_1$. Let

$$
\begin{aligned}
s_1(\beta_1, \cdots, \beta_p) &= \frac{1}{n}\sum_{j=1}^{p} \min\Big(\frac{|\beta_j|}{\tau_1}, 1\Big) A_{\cdot,j} + \lambda \sum_{j=1}^{p} \min\Big(\frac{|\beta_j|}{\tau_2}, 1\Big), \\
s_2(\beta_1, \cdots, \beta_p) &= -\frac{2}{n}\sum_{i=1}^{n} \min\Big(\frac{\sum_{j=1}^{p} A_{ij}|\beta_j|}{\tau_1}, 1\Big) + \frac{\alpha}{n}\sum_{j=1}^{p} \beta_j^2, \\
s_1^*(\beta_1) &= s_1(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|), \\
s_2^*(\beta_1) &= s_2(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|), \\
s^*(\beta_1) &= S(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|) = s_1^*(\beta_1) + s_2^*(\beta_1).
\end{aligned}
$$

Since $\frac{\partial s_1^*(\beta_1)}{\partial \beta_1} = 0$ and $\frac{\partial s_2^*(\beta_1)}{\partial \beta_1} > 0$ whenever $|\beta_1^*| > \tau_1$, we see that $|\beta_1^*|$ is not a local minimizer of $s^*(\beta_1)$, which is contrary to the assumption.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proof of Lemma 2. We prove by contradiction. We assume that $\boldsymbol{\beta}^* \neq \mathbf{0}$ is a local minimizer of $S(\boldsymbol{\beta})$ in (A.1) on $\mathbb{R}^p$. By construction of $S(\boldsymbol{\beta})$, we see that $|\boldsymbol{\beta}^*| = (|\beta_1^*|, \cdots, |\beta_p^*|)^T$ is also a local minimum of $S(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in R^p$. Without loss of generality, we assume that $|\beta_1^*| > 0$. Let

$$\begin{aligned}
s_1(\beta_1, \cdots, \beta_p) &= \frac{1}{n}\sum_{j=1}^{p} \min\Big(\frac{|\beta_j|}{\tau_{11}}, 1\Big) A_{\cdot,j} + \lambda \sum_{j=1}^{p} \min\Big(\frac{|\beta_j|}{\tau_2}, 1\Big), \\
s_2(\beta_1, \cdots, \beta_p) &= -\frac{2}{n}\sum_{i=1}^{n} \min\Big(\frac{\sum_{j=1}^{p} A_{ij}|\beta_j|}{\tau_{12}}, 1\Big) + \frac{\alpha}{n}\sum_{j=1}^{p} \beta_j^2, \\
s_1^*(\beta_1) &= s_1(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|), \\
s_2^*(\beta_1) &= s_2(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|), \\
s^*(\beta_1) &= S(\beta_1, |\beta_2^*|, \cdots, |\beta_p^*|) = s_1^*(\beta_1) + s_2^*(\beta_1).
\end{aligned}$$

We first consider the situation of $|\beta_1^*| = \tau_{11}$. Denote by the right derivative of $s_1^*(\beta_1)$ at $|\beta_1^*|$ to be b. By construction of $s_1^*(\cdot)$, its left derivative at $|\beta_1^*|$ must be $b + \frac{A_{\cdot,1}}{n\tau_{11}}$. Let c_1 and c_2 denote the left derivative and right derivative of $s_2^*(\beta_1)$ at $|\beta_1^*|$ respectively. Since $s^*(\beta_1)$ achieves a minimum at $|\beta_1^*|$, we have that $c_1 + b + \frac{A_{\cdot,1}}{n\tau_{11}} \leq 0$ and $c_2 + b \geq 0$, which implies that $c_2 - c_1 \geq \frac{A_{\cdot,1}}{n\tau_{11}}$. On the other hand, since $|\beta_1^*| > 0$, we have that c_1, $c_2 \in [-2\sum_{i=1}^{n} \frac{A_{i1}}{n\tau_{12}} + 2\frac{\alpha}{n}|\beta_1^*|, 2\frac{\alpha}{n}|\beta_1^*|]$, and thus $|c_2 - c_1| \leq \frac{2A_{\cdot,1}}{n\tau_{12}} \leq \frac{2A_{\cdot,1}}{2n\tau_{11}} = \frac{A_{\cdot,1}}{n\tau_{11}}$ because we have assumed that $\tau_{12} > 2\tau_{11}$, which is contrary to the fact that $c_2 - c_1 > \frac{A_{\cdot,1}}{n\tau_{11}}$.

Second, we consider the situation of $\tau_2 < |\beta_1^*| < \tau_{11}$. In this situation, the left derivative of $s_1^*(\beta_1)$ at $|\beta_1^*|$, b, is $\frac{A_{\cdot,1}}{n\tau_{11}}$, and the left derivative of $s_2^*(\beta_1)$ at $|\beta_1^*|$, c_1, belongs to $[-2\frac{A_{\cdot,1}}{n\tau_{12}} + 2\frac{\alpha}{n}|\beta_1^*|, 2\frac{\alpha}{n}|\beta_1^*|]$, which implies $b + c_1 > 0$ and is contrary to the the assumption of local minimum of $|\beta_1^*|$.

Third, we consider the situation of $0 < |\beta_1^*| \leq \tau_2$. We see that the left derivative of $s_1^*(\beta_1)$ at $|\beta_1^*|$, b, is $\frac{A_{\cdot,1}}{n\tau_{11}} + \frac{\lambda}{\tau_2}$, and the left derivative of $s_2^*(\beta_1)$ at $|\beta_1^*|$, c_1, belongs to $[-2\frac{A_{\cdot,1}}{n\tau_{12}} + 2\frac{\alpha}{n}|\beta_1^*|, 2\frac{\alpha}{n}|\beta_1^*|]$, which implies $b + c_1 > 0$ and is contrary to the the assumption of local minimum of $|\beta_1^*|$.

Finally, we consider the situation of $|\beta_1^*| > \tau_2$. Since $\frac{\partial s_1^*(\beta_1)}{\partial \beta_1} = 0$ and $\frac{\partial s_2^*(\beta_1)}{\partial \beta_1} > 0$ whenever $|\beta_1^*| > \tau_{12}$, we see that $|\beta_1^*|$ is not a local minimizer of $s^*(\beta_1)$, which is contrary to the assumption.

Other choices of the tuning parameters. This section focuses on situations involving different thresholding parameters for different approxima-

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tions of indicator functions in (2.3). Consider, for $\boldsymbol{\beta} \in [0, +\infty)^p$,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{j=1}^{p} \min\Big(\frac{\beta_j}{\tau_{11}}, 1\Big) A_{\cdot,j} - \frac{2}{n}\sum_{i=1}^{n} \min\Big(\frac{\sum_{j=1}^{p} A_{ij}\beta_j}{\tau_{12}}, 1\Big) \\ &\quad + \lambda \sum_{j=1}^{p} \min\Big(\frac{\beta_j}{\tau_2}, 1\Big) + \frac{\alpha}{n}\sum_{j=1}^{p} \beta_j^2, \end{aligned} \tag{A.1}$$

where τ_{11} and τ_{12} may not be equal.

First, we examine the cases of $\tau_{12} > 2\tau_{11}$ ($\tau_2 < \tau_{11}, \tau_{12}$).

LEMMA 2. *Let $\tau_{12} \geq 2\tau_{11}$ and $\tau_2 < \tau_{11}, \tau_{12}$. If there exists a local minimizer $\boldsymbol{\beta}^* \neq \mathbf{0}$ of $S(\boldsymbol{\beta})$ in (A.1), then $\beta_j^* = 0$ or $\tau_{11} < |\beta_j^*| \leq \tau_{12}$ for each $j \in \{1, \cdots, p\}$.*

Letting $\tau_{12} \geq 2\tau_{11}$, we have that in each iteration of Algorithm 1,

$$\begin{aligned} S^{(m)}(\boldsymbol{\beta}) &= \boldsymbol{\beta}^T \Big\{ \frac{\mathrm{diag}(A_{\cdot})\, I(\hat{\boldsymbol{\beta}}^{(m-1)} \leq \tau_{11})}{n\tau_{11}} + \lambda \frac{I(\hat{\boldsymbol{\beta}}^{(m-1)} \leq \tau_2)}{\tau_2} - \frac{2A_{\cdot}}{n\tau_{12}} \Big\} \\ &\quad + \frac{2}{n}\sum_{i=1}^{n} \max\Big(\frac{\sum_{j=1}^{p} A_{ij}\beta_j}{\tau_{12}} - 1, 0\Big) + \frac{\alpha}{n}\boldsymbol{\beta}^T\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in [0, +\infty)^p. \end{aligned} \tag{A.2}$$

It follows from (A.2) that once we have that $\hat{\boldsymbol{\beta}}^{(m-1)} = \mathbf{0}$ for some m, $S^{(m)}(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta} \in [0, +\infty)^p$, which terminates the DC iteration process, because $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} = \mathbf{0}$. This indicates that if $\tau_{12} \geq 2\tau_{11}$, the DC algorithm becomes sensitive to an initial value $\hat{\boldsymbol{\beta}}^{(0)}$.

Next, we examine the case of $0 < \tau_{12} < 2\tau_{11}$ ($\tau_2 < \tau_{11}, \tau_{12}$), where the DC algorithm is not sensitive as the first one. However, based on the results of a few numerical examples (not shown), we found that in this situation, even using one more parameter, the performance of finding the minimum cost subset did not improve over the proposed method.

Finally, we consider the case of $\tau_1 = \tau_{11} = \tau_{12}$ and $\tau_2 \geq \tau_1$. In this case, similar to Lemma 1, any local minimizer of $S(\boldsymbol{\beta})$ belongs to $[0, \tau_2]^p$, where the truncated L_1 penalty becomes a L_1 penalty that does not restrict the number of nonzero coordinates of a minimizer as an L_0 penalty does. In particular, in the situation with $\tau_1 = \tau_2$, $S(\boldsymbol{\beta})$ becomes a strictly convex function on $[0, \tau_2]^p$, which indicates that for any $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with $S(\boldsymbol{\beta}_1) = S(\boldsymbol{\beta}_2)$, $S(\frac{\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2}{2}) < S(\boldsymbol{\beta}_1)$. As a result, if there exists two minimum cost subsets B_1 and B_2 in the finite-sample situation, then by using $\tau_1 = \tau_2$, the corresponding method is more likely to select $B_1 \bigcup B_2$ as the minimum cost

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

subset.

The subgradient descent algorithm. For MCSS, we denote $\hat{\boldsymbol{\beta}}^{(m,1)} = (\hat{\beta}_1^{(m,1)}, \cdots, \hat{\beta}_n^{(m,1)})' = \hat{\boldsymbol{\beta}}^{(m-1)}$, use the following subgradient of $S^{(m)}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(m,t-1)}$:

$$\begin{aligned}
&\big(\nabla S^{(m)}(\hat{\boldsymbol{\beta}}^{(m,t-1)})\big) \\
=&\mathrm{diag}(\boldsymbol{A}_{\cdot})I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_1)/n\tau_1 + \lambda I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_2)/n\tau_2 - (1+\rho) * \boldsymbol{A}_{\cdot}^T/\tau_1 \\
&+ (1+\rho)\boldsymbol{A}^T I(\boldsymbol{A}\hat{\beta}^{(m,t-1)}/\tau_1 > 1)/n\tau_1 + 2\alpha\hat{\beta}^{(m,t-1)}/n
\end{aligned}$$

and then update $\hat{\boldsymbol{\beta}}^{(m,t)}$ until convergence to obtain $\hat{\boldsymbol{\beta}}^{(m)}$:

$$\hat{\boldsymbol{\beta}}^{(m,t)} = \hat{\boldsymbol{\beta}}^{(m,t-1)} - \frac{1}{2\sqrt{npt}}\nabla S^{(m)}(\hat{\boldsymbol{\beta}}^{(m,t-1)}). \tag{A.3}$$

For MCSS_ME, we denote $\hat{\boldsymbol{\beta}}^{(m,1)} = (\hat{\beta}_1^{(m,1)}, \cdots, \hat{\beta}_n^{(m,1)})' = \hat{\boldsymbol{\beta}}^{(m-1)}$, use the following subgradient of $S^{(m)}(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{(m,t-1)}$:

$$\begin{aligned}
&\big(\nabla S^{(m)}(\hat{\boldsymbol{\beta}}^{(m,t-1)})\big) \\
&= \big(\mathrm{diag}(\boldsymbol{A}_{\cdot})I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_1)/n\tau_1 + \lambda I(\hat{\boldsymbol{\beta}}^{(m-1)} \le \tau_2)/n\tau_2 - 2\boldsymbol{A}_{\cdot}/n\tau_1 \\
&- 2\gamma\boldsymbol{D}\hat{\boldsymbol{\beta}}^{(m-1)}/\tau_1^2 - 2\gamma\mathrm{diag}(I(\hat{\boldsymbol{\beta}}^{(m-1)} > \tau_1))\boldsymbol{D}\max(\hat{\boldsymbol{\beta}}^{(m-1)}/\tau_1 - 1, 0)/\tau_1\big) \\
&+ (1+\rho)\boldsymbol{A}^T I(\boldsymbol{A}\hat{\beta}^{(m,t-1)}/\tau_1 > 1)/n\tau_1 + 2\alpha\hat{\beta}^{(m,t-1)}/n \\
&+ 2\mathrm{diag}(\boldsymbol{C}_{\cdot})\hat{\beta}^{(m,t-1)}/n\tau_1 + 2\boldsymbol{C}\max(\hat{\beta}^{(m,t-1)}/\tau_1 - 1, 0)/n \\
&+ 2\mathrm{diag}(\boldsymbol{C}_{\cdot})\mathrm{diag}(I(\hat{\beta}^{(m,t-1)} > \tau_1))\max(\hat{\beta}^{(m,t-1)}/\tau 1 - 1, 0)/n\tau_1 \\
&+ 2\boldsymbol{C}\mathrm{diag}(\hat{\beta}^{(m,t-1)})I(\hat{\beta}^{(m,t-1)} > \tau_1)/n\tau_1
\end{aligned}$$

and then update $\hat{\boldsymbol{\beta}}^{(m,t)}$ by equation (A.3) until convergence to obtain $\hat{\boldsymbol{\beta}}^{(m)}$.

ACKNOWLEDGMENT

We are grateful to the editors and a reviewer for many constructive and helpful comments. This work was supported by NIH grants R01GM113250, R01HL105397 and R01HL116720, by NSF grants DMS-0906616 and DMS-1207771 and by NSFC grant 11571068. The authors thank Dr. Vandin for sharing the data.

REFERENCES

An, L. T. H. & Tao, P. D. (2005). The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research* **133**, 23–46.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S. *et al.* (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci.* **104**, 20007–20012.

Boca, S. M. (2010). Patient-oriented gene set analysis for cancer mutation data. *Genome Biol.* **11**, R112.

Brennan, C. W., Verhaak, R. G., McKenna, A., *et al.* (2013). The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477.

Ciriello, G., Cerami, E., Sander, C., Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406.

da Cunha Santos, G., Shepherd, F. A. and Tsao, M. S. (2011). EGFR mutations and lung cancer. *Annu Rev Pathol.* **6**, 49–69.

Masica, D. L., Karchin, R. (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.*, **71**, 4550-4561.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. *et al.* (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075.

Efroni, S. (2011). Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS One* **6**, e14437.

Feng, J., Kim, S. T., Liu, W., Kim, J. W., Zhang, Z., Zhu, Y., Berens, M., Sun, J., Xu, J. (2012). An integrated analysis of germline and somatic, genetic and epigenetic alterations at 9p21.3 in glioblastoma. *Cancer* **118**, 232–240.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. *et al.* (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucl. Acids Res.* **43 (D1)**, D805-D811.

Frattini, V., Trifonov, V., Chan, J.M., Castano, A., Lia, M., Abate, F., Keir, S.T., Ji, A.X., Zoppoli, P. *et al.* (2013). The integrated landscape of driver genomic alterations in glioblastoma. *Nature Genetics* **45**, 1141–1149.

Getz, G., Hofling, H., Mesirov, J. P., Golub, T. R., Meyerson, M., Tibshirani, R., Lander, E. S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500.

Gill, R. K., Yang, S. H., Meerzaman, D., Mechanic, L. E., Bowman, E. D., Jeon, H. S., Roy Chowdhur, S., Shakoori, A., Dracheva, T., Hong, K. M. *et al.* (2011). Frequent homozygous deletion of the LKB1/STK11 gene in non-small cell lung cancer. *Oncogene* **30**, 3784–3791.

Hahn, W. C. & Weinberg, R. A. (2002). Modelling the molecular circuitry of cancer. *Nat Rev Cancer* **2**, 331–341.

Hartmann, C., Bartels, G., Gehlhaar, C., Holtkamp, N., von Deimling, A. (2005). PIK3CA mutations in glioblastoma multiforme. *Acta Neuropathol.* **109**, 639–642.

Heinemann, V., Stintzing, S., Kirchner, T., Boeck, S., Jung, A. (2009). Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. *Cancer Treatment Reviews* **35**, 262–271.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A. *et al.* (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KANDOTH, C., MCLELLAN, M.D., VANDIN, F., YE, K., NIU, B., LU, C., XIE, M., ZHANG, Q., MCMICHAEL, J.F., WYCZALKOWSKI, M.A. *et al.* (2013). Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339.

LEISERSON, M. D. M., BLOKH, D., SHARAN, R., RAPHAEL, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* **9**, e1003054.

LI, C. & LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data, *Bioinformatics* **24**, 1175–1182.

LIU, L., LEI, J., WILLSEY, A. *et al.* (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism* **5**, 22.

LO, Y. L., HSIAO, C. F., JOU, Y. S., CHANG, G. C., TSAI, Y. H., SU, W. C., CHEN, Y. M., HUANG, M. S., CHEN, H. L., YANG, P. C *et al.* (2008). ATM polymorphisms and risk of lung cancer among never smokers., *Lung Cancer* **69**, 148–154.

MARDIS, E. R. & WILSON, R. K. (2009). Cancer genome sequencing: a review. *Hum Mol Genet* **18**, R163–R168.

MASICA, D. L. & KARCHIN, R. (2011). Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* **71**, 4550–4561.

MEYERSON, M., GABRIEL, S., GETZ, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685–696.

MILLER, C. A., SETTLE, S. H., SULMAN, E. P., ALDAPE, K. D., MILOSAVLJEVIC, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* **4**, 34.

QIU, Y. Q., ZHANG, S., ZHANG, X. S., CHEN, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *Bioinformatics* **11**, 26.

SCHAID, D.J., SINNWELL, J.P., JENKINS, G.D., MCDONNELL, S.K., INGLE, J.N., KUBO, M., GOSS, P.E., COSTANTINO, J.P., WICKERHAM, D.L. and WEINSHILBOUM, R.M. (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet Epidemiol* **36**, 3-16.

SCHWARTZENTRUBER, J., KORSHUNOV, A., LIU, X.Y., JONES, D.T., PFAFF, E., JACOB, K., STURM, D., FONTEBASSO, A.M., QUANG, D.A., TONJES, M., *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. (2012). *Nature* **482**, 226–231.

SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Am. Statist. Assoc.* **107**, 223–232.

SHOR, N.Z. (1985). *Minimization methods for non-diffrentiable functions.* Springer.

STARK, A. M., WITZEL, P., STREGE, R. J., HUGO, H. H., MEHDORN, H. M. (2003). p53, mdm2, EGFR, and msh2 expression in paired initial and recurrent glioblastoma multiforme. *J Neurol Neurosurg Psychiatry* **74**, 779–783.

STURM, D., BENDER, S., JONES, D.T., LICHTER, P., GRILL, J., BECHER, O., HAWKINS, C., MAJEWSKI, J., JONES, C., COSTELLO, J.F., IAVARONE, A. *et al.* (2014). Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat Rev Cancer.* **14**, 92–107.

THE CANCER GENOME ATLAS RESEARCH NETWORK (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature* **455**, 1061–1068.

THOMAS, R. K., BAKER, A. C., DEBIASI, R. M., WINCKLER, W., LAFRAMBOISE, T., LIN, W. M., WANG, M., FENG, W., ZANDER, T., MACCONAILL, L. *et al.* (2007). High-throughput oncogene mutation profiling in human cancer, *Nat Genet* **39**, 347–351.

bioRxiv preprint doi: https://doi.org/10.1101/117473; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Torkamani, A., Topo, E.J. and Schork, N.J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**, 265–272.

Turcan, S., Rohle, D., Goenka,A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W.M. *et al.* (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483.

Vandin, F., Upfal, E. and Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Research* **22**, 375–385.

Vogelstein, B. & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat Med* **10**, 789–799.

Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genome-wide association studies. *Am J Hum Genet* **81**, 1278-1283.

Yeang, C. H., McCormick, F. and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* **22**, 2605–2622.

Zhao, J., Zhang, S., Wu, L., Zhang, X. (2012). Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28**, 2940–2947.

Zhang, S. and Zhou, X.J (2014). Matrix factorization methods for integrative cancer genomics. *Methods Mol Biol.* **1176**, 229-242.

Zhuang, G., Song, W., Amato, K., Hwang, Y., Lee, K., Boothby, M., Ye, F., Guo, Y., Shyr, Y., Lin, L. *et al.* (2012). Effects of cancer-associated EPHA3 mutations on lung cancer. *J Natl Cancer Inst* **104**, 1182–1197.

Binghui Liu
School of Mathematics and Statistics
KLAS
Northeast Normal University
Changchun 130024, Jilin Province, China
E-mail: liubh100@nenu.edu.cn

Xiaotong Shen
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA
E-mail: xshen@stat.umn.edu

Chong Wu
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, MN 55455, USA
E-mail: wuxx0845@umn.edu

Wei Pan
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, MN 55455, USA
E-mail: weip@biostat.umn.edu